

Toward a Classification of Finite Partial-Monitoring Games[☆]

András Antos

*Machine Learning Group, Computer and Automation Research Institute of the Hungarian Academy of Sciences,
13-17 Kende utca, H-1111 Budapest, Hungary*

Gábor Bartók*, Dávid Pál, Csaba Szepesvári

Department of Computing Science, University of Alberta, Edmonton, Alberta, T6G 2E8, Canada

Abstract

Partial-monitoring games constitute a mathematical framework for sequential decision making problems with imperfect feedback: The learner repeatedly chooses an action, the opponent responds with an outcome, and then the learner suffers a loss and receives a feedback signal, both of which are fixed functions of the action and the outcome. The goal of the learner is to minimize his total cumulative loss. We make progress towards the classification of these games based on their minimax expected regret. Namely, we classify almost all games with two outcomes and a finite number of actions: We show that their minimax expected regret is either zero, $\tilde{\Theta}(\sqrt{T})$, $\Theta(T^{2/3})$, or $\Theta(T)$, and we give a simple and efficiently computable classification of these four classes of games. Our hope is that the result can serve as a stepping stone toward classifying all finite partial-monitoring games.

Keywords: Online algorithms, Online learning, Imperfect feedback, Regret analysis

1. Introduction

Partial-monitoring games constitute a mathematical framework for sequential decision making problems with imperfect feedback. They arise as a natural generalization of many sequential decision making problems with full or partial feedback such as learning with expert advice [2, 3, 4], the multi-armed bandit problem [5, 6, 7], label efficient prediction [8, 9], dynamic pricing [10, 11], the dark pool problem [12], the apple tasting problem [13], online convex optimization [14, 15], online linear [16] and convex optimization with bandit feedback [17].

A partial-monitoring game is a repeated game between two players: the *learner* and the *opponent*. In each round, the learner chooses an action and simultaneously the opponent chooses an outcome. Next, the learner receives a feedback signal and suffers a loss; however neither the loss nor the outcome are revealed to

[☆]Preliminary version of this paper appeared at ALT 2010, September 6–8, 2010, Canberra, Australia [1]. This work was supported in part by AICML, AITF (formerly iCore and AIF), NSERC and the PASCAL2 Network of Excellence under EC grant no. 216886.

*Corresponding authors

Email addresses: antos@cs.bme.hu (András Antos), bartok@cs.ualberta.ca (Gábor Bartók), dpal@cs.ualberta.ca (Dávid Pál), szepesva@cs.ualberta.ca (Csaba Szepesvári)

URL: <http://www.cs.bme.hu/~antos> (András Antos), <http://www.ualberta.ca/~bartok> (Gábor Bartók), <http://www.ualberta.ca/~dpal> (Dávid Pál), <http://www.ualberta.ca/~szepesva> (Csaba Szepesvári)

the learner. The feedback and the loss are fixed functions of the action and the outcome, and these functions are known by both players. The main feature of this model is that it captures that the learner has imperfect or partial information about the outcome sequence. In this work, we make the natural assumption that the opponent is *oblivious*, that is, the opponent does not have access to the learner's actions.

The goal of the learner is to keep his cumulative loss small. However, since the opponent could choose the outcome sequence so that the learner suffers as high loss as possible, it is too much to ask for an absolute guarantee for the cumulative loss. Instead, a competitive viewpoint is taken and the cumulative loss of the learner is compared with the cumulative loss of the best among all the constant strategies, i.e., strategies that choose the same action in every round. The difference between the cumulative loss of the learner and the cumulative loss of the best constant strategy is called the *regret*.

Generally, the regret grows with the number of rounds of the game. If the growth is sublinear then the learner is said to be Hannan consistent¹, and in the long run the learner's average loss per round approaches the average loss per round of the best action.

Designing learning algorithms with low regret is the main focus of study of partial-monitoring games. For a given game, the ultimate goal is to find out its optimal worst-case (minimax) regret, and design an algorithm that achieves it. The minimax regret can be viewed as an inherent measure of how hard the game is for the learner. The motivation behind this paper was the desire to determine the minimax regret and design an algorithm achieving it for each game in a large class.

In this paper we restrict our attention to games with a finite number of actions and *two outcomes*. This class is a subset of the class of *finite partial-monitoring games*, introduced by Piccolboni and Schindelhauer [19], in which both the set of actions and the set of outcomes are finite.

1.1. Previous Results

For full-information games (i.e., when the feedback determines the outcome) with N actions and losses lying in the interval $[0, 1]$, there exists a randomized algorithm with expected regret at most $\sqrt{T \ln(N)/2}$ where T is the time horizon (see e.g., Lugosi and Cesa-Bianchi [20, Chapter 4] and references therein). Furthermore, it is known that this upper bound is tight: There exist full-information games with losses lying in the interval $[0, 1]$ for which the worst-case expected regret of any algorithm is at least $\Omega(\sqrt{T \ln N})$ [20, Chapter 3].

Another special case of partial-monitoring games is the multi-armed bandit game, where the learner's feedback is the loss of the action he chooses. For a multi-armed bandit game with N actions and losses lying in the interval $[0, 1]$, the INF algorithm [21] has expected regret at most $O(\sqrt{TN})$. (The well-known Exp3 algorithm [5] achieves the bound $O(\sqrt{TN \log N})$.) It is also known that the bound $O(\sqrt{TN})$ is optimal [5].

Piccolboni and Schindelhauer [19] introduced finite partial-monitoring games. They showed that, for any finite game, either there is a strategy for the learner that achieves regret of at most $O(T^{3/4}(\ln T)^{1/2})$ or the worst-case expected regret of any learner is $\Omega(T)$. Cesa-Bianchi et al. [22] improved this result and showed that Piccolboni and Schindelhauer's algorithm achieves $O(T^{2/3})$ regret. They also gave an example of a game with worst-case expected regret at least $\Omega(T^{2/3})$. More recently, Lugosi et al. [23] designed algorithms and proved upper bounds in a slightly different setting, where the feedback signal is a possibly noisy function of the outcome or both the action and the outcome.

However, from these results it is unclear what determines which games have minimax regret $\Theta(\sqrt{T})$, which games have minimax regret $\Theta(T^{2/3})$ and whether there exist finite games with minimax regret not

¹Hannan consistency is named after James Hannan who was the first to design a learning algorithm with sublinear regret for finite games with full feedback [18].

belonging to either of these categories. Cesa-Bianchi et al. [22] note that: “It remains a challenging problem to characterize the class of problems that admit rates of convergence faster than $O(n^{-1/3})$.”²

1.2. Our Results

We classify the minimax expected regret of finite partial-monitoring games with *two outcomes*. From our classification we exclude certain “degenerate games”; their precise definition is given later in the paper. We show that the minimax regret of any non-degenerate game falls into one of the four categories: 0, $\widetilde{O}(\sqrt{T})$, $\Theta(T^{2/3})$, $\Theta(T)$ and no other option is possible³. We call the four classes of games *trivial*, *easy*, *hard*, and *hopeless*, respectively. We give a simple and efficiently computable geometric characterization of these four classes.

Additionally, we show that each of the four classes admits a computationally efficient learning algorithm achieving the minimax expected regret, up to logarithmic factors. In particular, we design an efficient learning algorithm for easy games with expected regret at most $\widetilde{O}(\sqrt{T})$. For hard games, the algorithm of Cesa-Bianchi et al. [22] has $O(T^{2/3})$ regret. For trivial games, a simple algorithm that chooses the same action in every round has zero regret. For hopeless games, any algorithm has $\Theta(T)$ regret.

2. Basic Definitions and Notations

A finite partial-monitoring game is specified by a pair of $N \times M$ matrices (\mathbf{L}, \mathbf{H}) where N is the number of actions, M is the number of outcomes, \mathbf{L} is the *loss matrix*, and \mathbf{H} is the *feedback matrix*. We use the notation $\underline{n} = \{1, \dots, n\}$ for any integer and denote the actions and outcomes by integers starting from 1, so the action set is \underline{N} and the outcome set is \underline{M} . We denote by $\ell_{i,j}$ and $h_{i,j}$ ($i \in \underline{N}$, $j \in \underline{M}$) the entries of \mathbf{L} and \mathbf{H} , respectively. We denote by ℓ_i the i -th row ($i \in \underline{N}$) of \mathbf{L} , and we call it the *loss vector of action i* . The elements of \mathbf{L} are arbitrary real numbers. The elements of \mathbf{H} belong to some alphabet Σ , we only assume that the learner is able to distinguish two different elements of the alphabet. We often use the set of natural or real numbers as the alphabet.

The matrices \mathbf{L}, \mathbf{H} are known by both the learner and the opponent. The game proceeds in T rounds. In each round $t = 1, 2, \dots, T$, the learner chooses an action $I_t \in \underline{N}$ and simultaneously the opponent chooses an outcome $J_t \in \underline{M}$, then the learner receives the feedback h_{I_t, J_t} . Nothing else is revealed to the learner; in particular J_t and the loss ℓ_{I_t, J_t} remain hidden.

In principle, both I_t and J_t can be chosen randomly. However, to simplify our treatment, we assume that the opponent is deterministic and oblivious to the actions of the learner. Equivalently, we can assume that the sequence of outcomes J_1, J_2, \dots, J_T is a fixed deterministic sequence chosen before the first round of the game. On the other hand, it is important to allow the learner to choose his actions I_t randomly. A randomized strategy (algorithm) A of the learner is a sequence of random functions I_1, I_2, \dots, I_T where each of the functions maps the feedback from the past outcomes (and learner’s internal random “bits”) to an action; formally $I_t : \Sigma^{t-1} \times \Omega \rightarrow \underline{N}$.

The learner is scored according to the loss matrix. In each round t , the learner incurs *instantaneous loss* ℓ_{I_t, J_t} . The goal of the learner is to keep his *cumulative loss* $\sum_{t=1}^T \ell_{I_t, J_t}$ small. The (*cumulative*) *regret* of an algorithm A is defined as

$$\widehat{R}_T = \widehat{R}_T(A, G) = \sum_{t=1}^T \ell_{I_t, J_t} - \min_{i \in \underline{N}} \sum_{t=1}^T \ell_{i, J_t}.$$

²They used n instead of T and by rate they mean the average regret per time step.

³The notation \widetilde{O} and $\widetilde{\Theta}$ hides poly-logarithmic factors in T .

In other words, the regret is the excess loss of the learner compared to the loss of the best constant action. We denote by $R_T = R_T(A, G) = \mathbf{E}[R_T(A, G)]$ the (cumulative) expected regret. Let the worst-case expected regret of A when used in $G = (\mathbf{L}, \mathbf{H})$ be

$$\bar{R}_T(A, G) = \sup_{J_{1:T} \in \underline{M}^T} R_T(A, G),$$

where the supremum is taken over all outcome sequences $J_{1:T} = (J_1, J_2, \dots, J_T) \in \underline{M}^T$. The minimax expected regret of G (or minimax regret, for short) is:

$$R_T(G) = \inf_A \bar{R}_T(A, G) = \inf_A \sup_{J_{1:T} \in \underline{M}^T} R_T(A, G),$$

where the infimum is taken over all randomized strategies A . Note that, since $R_T(A, G) \geq 0$ for constant outcome sequences, $R_T(G) \geq 0$ also holds.

We identify the set of all probability distributions over the set of outcomes \underline{M} with the probability simplex $\Delta_M = \{p \in \mathbb{R}^M : \sum_{j=1}^M p(j) = 1, \forall j \in \underline{M}, p(j) \geq 0\}$. We use $\langle \cdot, \cdot \rangle$ to denote the standard dot product.

3. Characterization of Games with Two Outcomes

In this section, we formally phrase our main characterization result. We need a preliminary definition that is useful for any finite game:

Definition 1 (Properties of Actions). Let $G = (\mathbf{L}, \mathbf{H})$ be a finite partial-monitoring game with N actions and M outcomes. Let $i \in \underline{N}$ be one of its actions.

- Action i is called *dominated* if for any $p \in \Delta_M$ there exists an action i' such that $\ell_{i'} \neq \ell_i$ and $\langle \ell_{i'}, p \rangle \leq \langle \ell_i, p \rangle$.
- Action i is called *non-dominated* if it is not dominated.
- Action i is called *degenerate* if it is dominated and there exists a distribution $p \in \Delta_M$ such that for all $i' \in \underline{N}$, $\langle \ell_i, p \rangle \leq \langle \ell_{i'}, p \rangle$.
- Action i is called *all-revealing* if any pair of outcomes j, j' , $j \neq j'$ satisfies $h_{i,j} \neq h_{i,j'}$.
- Action i is called *none-revealing* if any pair of outcomes j, j' satisfies $h_{i,j} = h_{i,j'}$.
- Action i is called *partially-revealing* if it is neither all-revealing nor none-revealing.
- All-revealing and partially-revealing actions together are called *revealing* actions.
- Two or more actions with the same loss vector are called *duplicate* actions.

The property of being dominated has an equivalent dual definition. Namely, action i is dominated if there exists a set of actions with loss vectors not equal to ℓ_i such that some convex combination of their loss vectors is componentwise upper bounded by ℓ_i .

In games with $M = 2$ outcomes, each action is either all-revealing or none-revealing. This dichotomy is one of the key properties that lead to the classification theorem for two-outcome games. To emphasize the

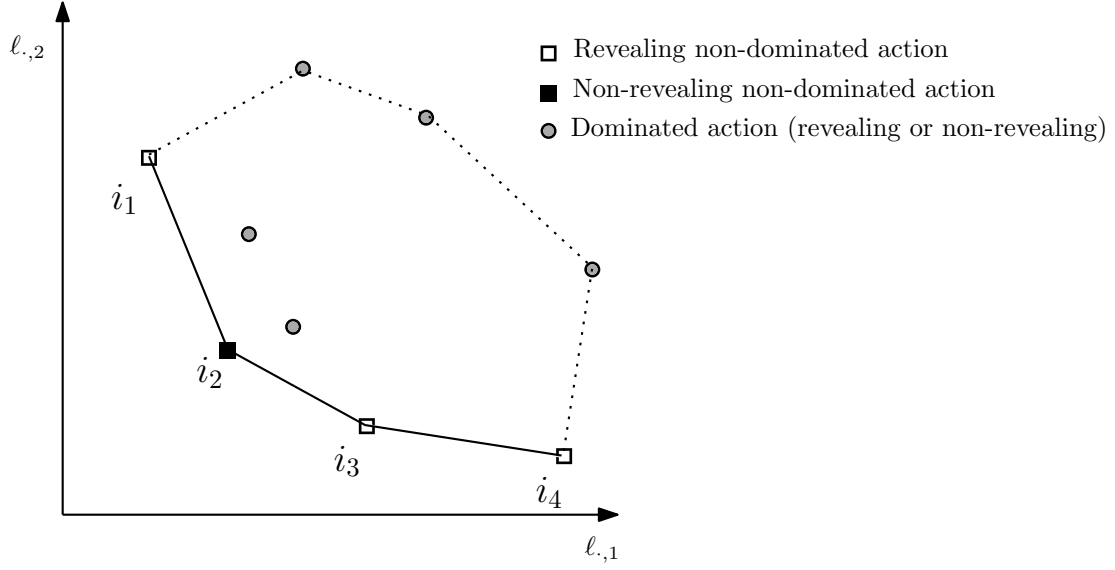


Figure 1: The figure shows each action i as a point in \mathbb{R}^2 with coordinates $(\ell_{i,1}, \ell_{i,2})$. The solid line connects the chain of non-dominated actions, which, by convention are ordered according to their loss for the first outcome.

dichotomy, from now on we will refer to them as revealing and *non-revealing* whenever it is clear from the context that $M = 2$.

The above property also allows us to assume without loss of generality that there are no duplicate actions. Clearly, if multiple actions with the same loss vector exist, all but one can be removed (together with the corresponding rows of \mathbf{L} and \mathbf{H}) without changing the minimax regret: If all of them are non-revealing, we keep one of the actions and remove all the others. Otherwise, we keep a revealing action and remove the others. Then replacing any algorithm by one that, instead of a removed action, chooses always the corresponding kept action, its loss cannot increase and equals to the loss of this algorithm for the original game. So the two games have the same minimax regret.

The concepts of dominated and non-dominated actions can be visualized for two-outcome games by drawing the loss vector of each action as a point in \mathbb{R}^2 . The points corresponding to the non-dominated actions lie on the bottom-left boundary of the convex hull of the set of all the actions, as shown in Figure 1. Enumerating the non-dominated actions ordered according to their loss for the first outcome gives rise to a sequence (i_1, i_2, \dots, i_K) , which we call the *chain of non-dominated actions*.

To state the classification theorem, we introduce the following conditions.

Separation Condition. A two-outcome game G satisfies the separation condition if, after removing duplicate actions, its chain of non-dominated actions does **not** have a pair of consecutive actions i_k, i_{k+1} such that both of them are non-revealing. The set of games satisfying this condition will be denoted by \mathcal{S} .

Non-degeneracy Condition. A two-outcome game G is degenerate if it has a degenerate revealing action. If G is not degenerate, we call it non-degenerate and we say that it satisfies the non-degeneracy condition.

As we will soon see, the separation condition is the key to distinguish between *hard* and *easy* games. On the other hand, the non-degeneracy condition is merely a technical condition that we need in our proofs. The set of degenerate games is excluded from the characterization, as we do not know the minimax regret of these games. We are now ready to state our main result.

Theorem 2 (Classification of Two-Outcome Partial-Monitoring Games). *Let \mathcal{S} be the set of all finite partial-monitoring games with two outcomes that satisfy the separation condition. Let $G = (\mathbf{L}, \mathbf{H})$ be a game with two outcomes that satisfies the non-degeneracy condition. Let K be the number of non-dominated actions in G , counting duplicate actions only once. The minimax expected regret $R_T(G)$ satisfies*

$$R_T(G) = \begin{cases} 0 & (\forall T), & K = 1; & (1a) \\ \widetilde{\Theta}(\sqrt{T}), & K \geq 2, G \in \mathcal{S}; & (1b) \\ \Theta(T^{2/3}), & K \geq 2, G \notin \mathcal{S}, G \text{ has a revealing action}; & (1c) \\ \Theta(T), & \text{otherwise.} & (1d) \end{cases}$$

We call the games in cases (1a)–(1d) *trivial*, *easy*, *hard*, and *hopeless*, respectively. Case (1a) is proven by the following lemma which shows that a trivial game is also characterized by having 0 minimax regret in a single round or by having an action “dominating” alone all the others:

Lemma 3. *For any finite partial-monitoring game, the following four statements are equivalent:*

- a) *The minimax regret is zero for each T .*
- b) *The minimax regret is zero for some T .*
- c) *There exists a (non-dominated) action $i \in \underline{N}$ whose loss is not larger than the loss of any other action irrespectively of the choice of Nature’s action.*
- d) *The game is trivial, i.e., $K = 1$ (using the definition in Theorem 2).*

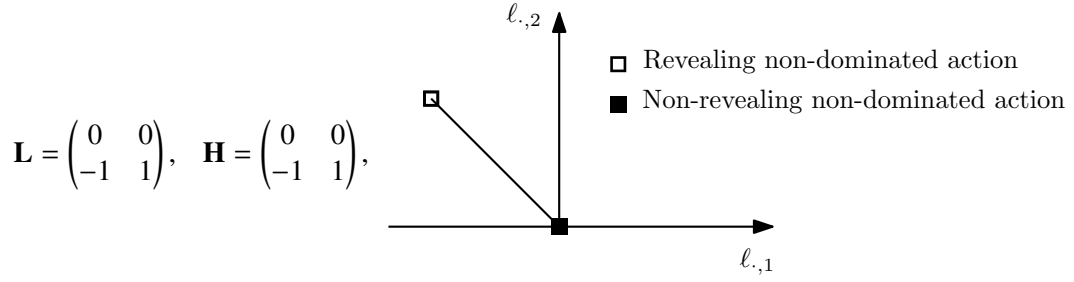
The proof of this lemma can be found in the Appendix. Case (1d) of Theorem 2 is proven in the Appendix as well. The upper bound of case (1c) can be derived from a result of Cesa-Bianchi et al. [22]: Recall that the entries of \mathbf{H} can be changed without changing the information revealed to the learner as long as one does not change the pattern of which elements in a row are equal and different. Cesa-Bianchi et al. [22] show that if the entries of \mathbf{H} can be chosen such that $\text{rank}(\mathbf{H}) = \text{rank}\begin{pmatrix} \mathbf{H} \\ \mathbf{L} \end{pmatrix}$ then $O(T^{2/3})$ expected regret is achievable. This condition holds trivially for two-outcome games with at least one revealing action and $N \geq 2$. It remains to prove the upper bound for case (1b), the lower bound for (1b), and the lower bound for (1c); we prove these in Sections 5, 6, and 7, respectively.

4. Examples

Before we dive into the proof of Theorem 2, we give a few examples of finite partial-monitoring games with two outcomes and show how the theorem can be applied. For each example we present the matrices \mathbf{L}, \mathbf{H} and depict the loss vectors of actions as points in \mathbb{R}^2 .

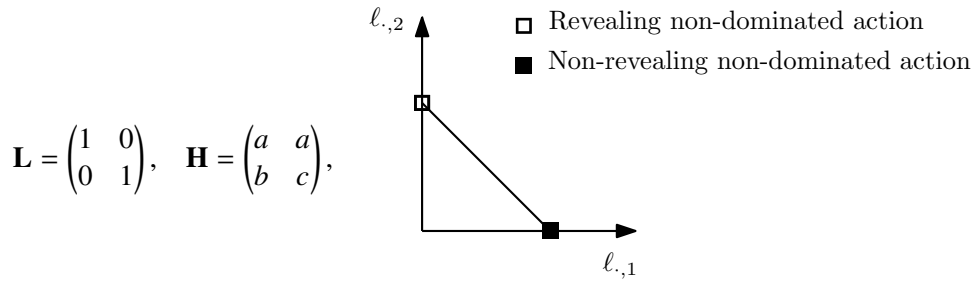
Example 4 (One-Armed Bandit). We start with an example of a multi-armed bandit game. Multi-armed bandit games are those where the feedback equals the instantaneous loss, that is, when $\mathbf{L} = \mathbf{H}$.⁴

⁴“Classically”, non-stochastic multi-armed bandit problems are defined by the restriction that in no round Learner can gain any information about the losses of actions other than the chosen one, that is, \mathbf{L} is not known in advance to Learner. (Also, the domain set of losses is often infinite there ($M = \infty$).) When $\mathbf{H} = \mathbf{L}$ in our setting, depending on \mathbf{L} , this might or might not be the case; the “classical bandit” problem with losses constrained to a finite set is a special case of games with $\mathbf{H} = \mathbf{L}$, however, the latter condition allows also other types of games where the Learner can recover the losses of actions not chosen, and so which could be “easier” than classical bandits due to the knowledge of \mathbf{L} . Nevertheless, it is easy to see that these games are *at most* as hard as classical bandit games.



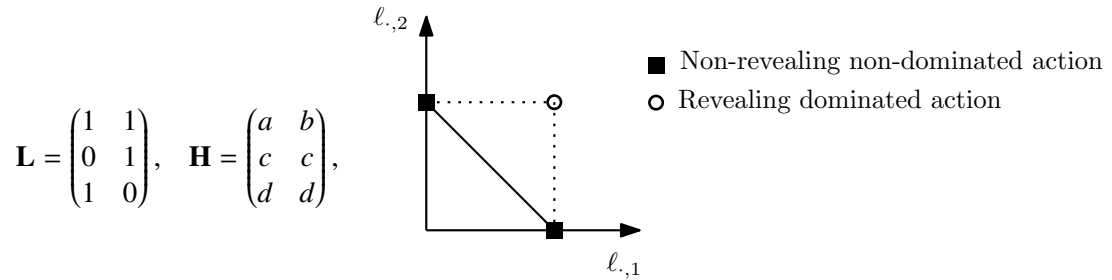
Because the loss of the first action is 0 regardless of the outcome, and the loss varies only for the second action, we call this game a *one-armed bandit* game. Both actions are non-dominated and the second one is revealing, therefore it is an easy game and according to Theorem 2 its minimax regret is $\tilde{\Theta}(\sqrt{T})$. (For this specific game, it can be shown that it is in fact $\Theta(\sqrt{T})$.)

Example 5 (Apple Tasting). Consider an orchard that wants to hand out its crop of apples for sale. However, some of the apples might be rotten. The orchard can do a sequential test. Each apple can be either tasted (which reveals whether the apple is healthy or rotten) or the apple can be given out for sale. If a rotten apple is given out for sale, the orchard suffers a unit loss. On the other hand, if a healthy apple is tasted, it cannot be sold and, again, the orchard suffers a unit loss. This can be formalized by the following partial-monitoring game [13]:



The first action corresponds to giving out the apple for sale, the second corresponds to tasting the apple; the first outcome corresponds to a rotten apple, the second outcome corresponds to a healthy apple. Both actions are non-dominated and the second one is revealing, therefore it is an easy game and according to Theorem 2 the minimax regret is $\tilde{\Theta}(\sqrt{T})$. This is apparently a new result for this game. Also notice that the picture is a just a translation of the picture for the one-armed bandit.

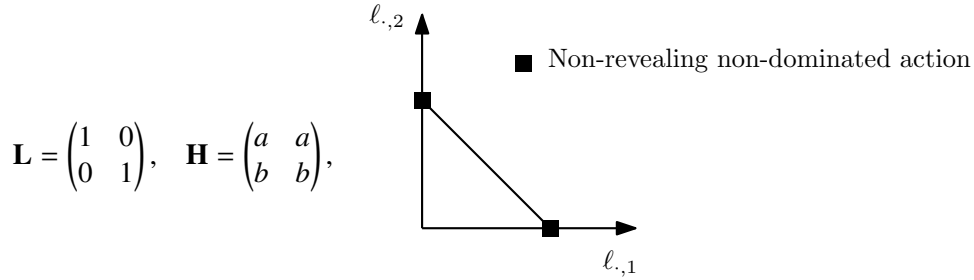
Example 6 (Label Efficient Prediction). Consider a situation when we would like to sequentially classify emails as spam or as legitimate. For each email we have to output a prediction, and additionally we can request, as feedback, the correct label from the user. If we classify an email incorrectly or we request its label, we suffer a unit loss. (If the email is classified correctly and we do not request the feedback, no loss is suffered.) This can be formalized by the following partial-monitoring game [22]:



where the first action corresponds to a label request, and the second and the third action correspond to a prediction (spam and legitimate, respectively) without a request. The outcomes correspond to spam and legitimate emails.

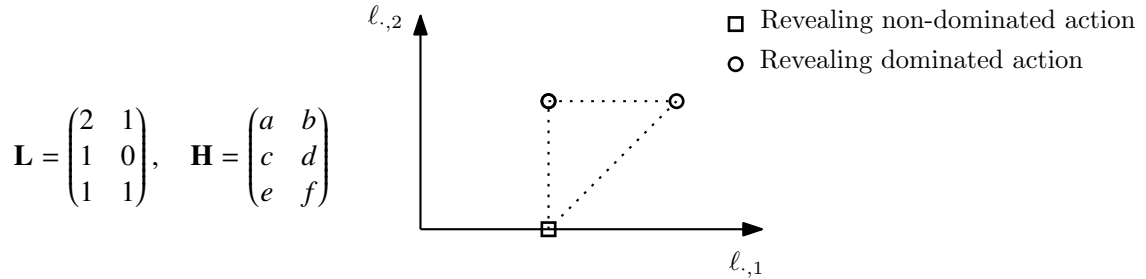
We see that the chain of non-dominated actions contains two neighboring non-revealing actions and there is a dominated revealing action. Therefore, it is a hard game and, by Theorem 2, the minimax regret is $\Theta(T^{2/3})$. This specific example was the only game known so far with minimax regret at least $\Omega(T^{2/3})$ [22, Theorem 5.1].

Example 7 (A Hopeless Game). The following game is an example where the feedback does not reveal any information about the outcome:



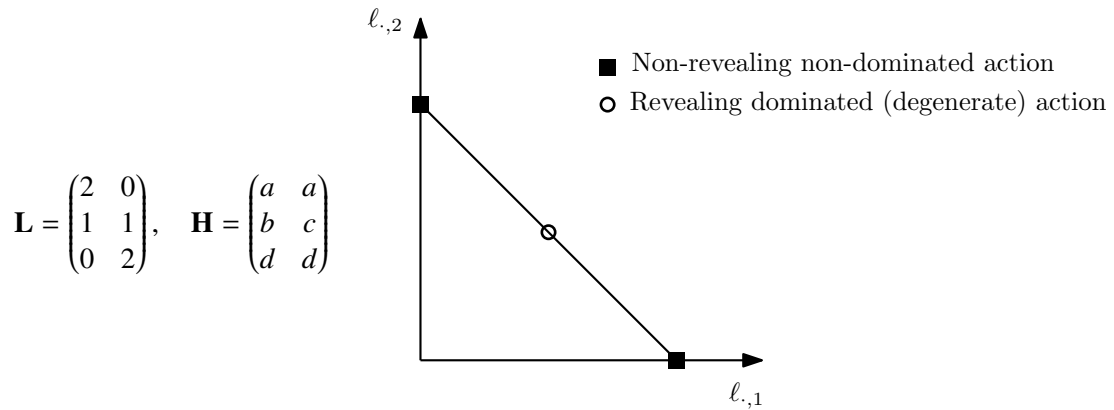
Because both actions are non-revealing and non-dominated, it is a hopeless game and thus its minimax regret is $\Theta(T)$.

Example 8 (A Trivial Game). In the following game, the best action, regardless of the outcome sequence, is action 2. A learner that chooses this action in every round is guaranteed to have zero regret.



Because this game has only one non-dominated action (action 2), it is a trivial game and thus its minimax regret is 0.

Example 9 (A Degenerate Game). The next game does not satisfy the non-degeneracy condition and therefore Theorem 2 does not apply.



Its minimax regret is between $\Omega(\sqrt{T})$ and $O(T^{2/3})$. It remains an open problem to close this gap and determine the exact rate of growth.

5. Upper bound for easy games

In this section we present our algorithm for games satisfying the separation condition and the non-degeneracy condition, and prove that it achieves $\tilde{O}(\sqrt{T})$ regret with high probability. We call the algorithm APPLE TREE since it builds a binary tree, leaves of which are apple tasting games.

5.1. The algorithm

In the first step of the algorithm we can purify the game by first removing the dominated actions and then the duplicates as mentioned beforehand.

The idea of the algorithm is to recursively split the game until we arrive at games with two actions only. Now, if one has only two actions in a partial-information game, the game must be either a full-information game (if both actions are revealing) or an instance of a one-armed bandit (with one revealing and one non-revealing action).

To see why this latter case corresponds to one-armed bandits, assume without loss of generality that the first action is the revealing action. Now, it is easy to see that the regret of a sequence of actions in a game does not change if the loss matrix is changed by subtracting the same number from a column.⁵ By subtracting $\ell_{2,1}$ from the first and $\ell_{2,2}$ from the second column we thus get the equivalent game where the second row of the loss matrix is zero, arriving at a one-armed bandit game (see Example 4). Since a one-armed bandit is a special form of a two-armed bandit, one can use Exp3.P due to Auer et al. [5] to achieve the $O(\sqrt{T})$ regret.

Now, if there are more than two actions in the game, then the game is split, putting the first half of the actions into the first and the second half into the second subgame, with a *single common shared action*. Recall that, in the chain of non-dominated actions, the actions are ordered according to their losses corresponding to the *first* outcome. This is continued until the split results in games with two actions only. The recursive splitting of the game results in a binary tree (see Figure 2). The idea of the strategy played at an internal node of the tree is as follows: An outcome sequence of length T determines the frequency ρ_T of outcome 2. If this frequency is small, the optimal action is one of the actions of G_1 , the first subgame (simply because then the frequency of outcome 1 is high and G_1 contains the actions with the smallest loss for the first outcome). Conversely, if this frequency is large, the optimal action is one of the actions of G_2 . In some intermediate range, the optimal action is the action shared between the subgames. Let the boundaries of this range be $\rho_1^* < \rho_2^*$ (ρ_1^* is thus the solution to $(1 - \rho)\ell_{s-1,1} + \rho\ell_{s-1,2} = (1 - \rho)\ell_{s,1} + \rho\ell_{s,2}$ and ρ_2^* is the solution to $(1 - \rho)\ell_{s+1,1} + \rho\ell_{s+1,2} = (1 - \rho)\ell_{s,1} + \rho\ell_{s,2}$, where $s = \lceil K/2 \rceil$ is the index of the action shared between the two subgames.)

If we knew ρ_T , a good solution would be to play a strategy where the actions are restricted to that of either game G_1 or G_2 , depending on whether $\rho_T \leq \rho_1^*$ or $\rho_T \geq \rho_2^*$. (When $\rho_1^* \leq \rho_T \leq \rho_2^*$ then it does not matter which action-set we restrict the play to, since the optimal action in this case is included in both sets.) There are two difficulties. First, since the outcome sequence is not known in advance, the best we can hope for is to know the running frequencies $\rho_t = \frac{1}{t} \sum_{s=1}^t \mathbb{I}(J_s = 2)$. However, since the game is a partial-information game, the outcomes are not revealed in all time steps, hence, even ρ_t is inaccessible.

⁵As a result, for any algorithm, if R_T is its regret at time T when measured in the game with the modified loss matrix, the algorithm's "true" regret will also be R_T (i.e., the algorithm's regret when measured in the original, unmodified game). Piccolboni and Schindelhauer [19] exploit this idea, too.

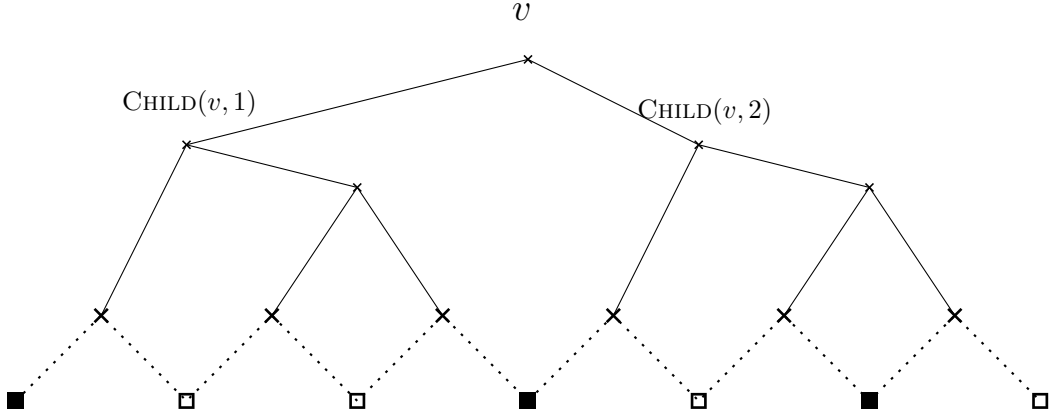


Figure 2: The binary tree built by the algorithm. The leaf nodes represent neighboring action pairs.

Nevertheless, for now let us assume that ρ_t was available. Then one idea would be to play a strategy restricted to the actions of either game G_1 or G_2 as long as ρ_t stays below ρ_1^* or above ρ_2^* . Further, when ρ_t becomes larger than ρ_2^* while previously the strategy played the action of G_1 then we have to switch to the game G_2 . In this case, we start a fresh copy (a *reset*) of a strategy playing in G_2 . The same happens when a switch from G_2 to game G_1 is necessary. These resets are necessary because at the leaves we play according to strategies that use weights that depend on the cumulated losses of the actions *exponentially*. To see an example when without resets the algorithm fails to achieve a small regret consider the case when there are 3 actions, the middle one being revealing. Assume that during the first $T/2$ time steps the frequency of outcome 2 oscillates between the two boundaries so that the algorithm switches constantly back and forth between the games G_1 and G_2 . Assume further that in the second half of the game, the outcome is always 2. This way the optimal action will be 3. Nevertheless, up to time step $T/2$, the player of G_2 will only see outcome 1 and thus will think that action 2 is the optimal action. In the second half of the game, he will not have enough time to recover and will play action 2 for too long. Resetting the algorithms of the subgames avoids this behavior.

If the number of switches was large, the repeated resetting of the strategies could be equally problematic. Luckily this cannot happen, hence the resetting does minimal harm. We will in fact show that this generalizes to the case even when ρ_t is estimated based on partial feedback (see Lemma 11).

Let us now turn to how ρ_t is estimated. As mentioned in Section 3, mapping a row of \mathbf{H} bijectively leads to an equivalent game, thus for $M = 2$ we can assume without loss of generality that in any round, the algorithm receives (possibly random) feedback $H_t \in \{1, 2, *\}$: if a revealing action is played in the round, $H_t = J_t \in \{1, 2\}$, otherwise $H_t = *$. Let $\mathcal{H}_{1:t-1} = (I_1, H_1, \dots, I_{t-1}, H_{t-1}) \in (\underline{N} \times \Sigma)^{t-1}$, the (random) history of actions and observations up to time step $t - 1$. If the algorithm choosing the actions decides with probability $p_t \in (0, 1]$ to play a revealing action (p_t can depend on $\mathcal{H}_{1:t-1}$) then $\mathbb{I}(H_t = 2) / p_t$ is a simple unbiased estimate of $\mathbb{I}(J_t = 2)$ (in fact, $\mathbf{E}[\mathbb{I}(H_t = 2) / p_t | \mathcal{H}_{1:t-1}] = \mathbb{I}(J_t = 2)$). As long as p_t does not drop to a too low value, $\hat{\rho}_t = \frac{1}{t} \sum_{s=1}^t \frac{\mathbb{I}(H_s=2)}{p_s}$ will be a relatively reliable estimate of ρ_t (see Lemma 12). However reliable this estimate is, it can still differ from ρ_t . For this reason, we push the boundaries determining game switches towards each other:

$$\rho'_1 = \frac{2\rho_1^* + \rho_2^*}{3}, \quad \rho'_2 = \frac{\rho_1^* + 2\rho_2^*}{3}. \quad (2)$$

We call the resulting algorithm APPLE TREE, because the elementary partial-information 2-action games in the bottom essentially correspond to instances of the apple tasting problem (see Example 5). The algo-

function MAIN(G, T, δ)

Input: $G = (\mathbf{L}, \mathbf{H})$ is a game, T is a horizon,
 $0 < \delta < 1$ is a confidence parameter

```

1:  $G \leftarrow \text{PURIFY}(G)$ 
2: BUILDTREE(root,  $G, \delta$ )
3: for  $t \leftarrow 1$  to  $T$  do
4:   PLAY(root)
5: end for
```

Figure 3: The main entry point of the APPLETREE algorithm

function INITETA(G, T)

Input: G is a game, T is a horizon

```

1: if ISREVEALING( $G, 2$ ) then
2:    $\eta(v) \leftarrow \sqrt{8 \ln 2 / T}$ 
3: else
4:    $\eta(v) \leftarrow \gamma(v)/4$ 
5: end if
```

Figure 4: The initialization routine INITETA.

function BUILDTREE(v, G, δ)

Input: $G = (\mathbf{L}, \mathbf{H})$ is a game, v is a tree node

```

1: if NUMOFACTIONS( $G$ ) = 2 then
2:   if not ISREVEALING( $G, 1$ ) then
3:      $G \leftarrow \text{SWAPACTIONS}(G)$ 
4:   end if
5:    $w_i(v) \leftarrow 1/2, i = 1, 2$ 
6:    $\beta(v) \leftarrow \sqrt{\ln(2/\delta)/(2T)}$ 
7:    $\gamma(v) \leftarrow 8\beta(v)/(3 + \beta(v))$ 
8:   INITETA( $G, T$ )
9: else
10:  ( $G_1, G_2$ )  $\leftarrow \text{SPLITGAME}(G)$ 
11:  BUILDTREE(CHILD( $v, 1$ ),  $G_1, \delta/(4T)$ )
12:  BUILDTREE(CHILD( $v, 2$ ),  $G_2, \delta/(4T)$ )
13:   $g(v) \leftarrow 1, \hat{p}(v) \leftarrow 0, t(v) \leftarrow 1$ 
14:  ( $\rho'_1(v), \rho'_2(v)$ )  $\leftarrow \text{BOUNDARIES}(G)$ 
15: end if
16:  $G(v) \leftarrow G$ 
```

Figure 5: The tree building procedure

algorithm's main entry point is shown on Figure 3. Its inputs are the game $G = (\mathbf{L}, \mathbf{H})$, the time horizon and a confidence parameter $0 < \delta < 1$. The algorithm first eliminates the dominated and duplicate actions. This is followed by building a tree, which is used to store variables necessary to play in the subgames (Figure 5): If the number of actions is 2, the procedure initializes various parameters that are used either by a bandit algorithm (based on Exp3.P [5]), or by the exponentially weighted average algorithm (EWA) [4]. In the other case, it calls itself recursively on the split subgames and with an appropriately decreased confidence parameter.

The main worker routine is called **PLAY**. This is again a recursive function (see Figure 6). The special case when the number of actions is two is handled in routine **PLAYATLEAF**, which will be discussed later. When the number of actions is larger, the algorithm recurses to play in the subgame that was remembered as the game to be preferred from the last round and then updates its estimate of the frequency of outcome 2 based on the information received. When this estimate changes so that a switch of the current preferred game is necessary, the algorithm resets the algorithms in the subtree corresponding to the game switched to, and changes the variable storing the index of the preferred game. The **RESET** function used for this purpose, shown on Figure 7, is also recursive.

At the leaves, when there are only two actions, either EWA or Exp3.P is used. These algorithms are used with their standard optimized parameters (see Corollary 4.2 for the tuning of EWA, and Theorem 6.10 for the tuning of Exp3.P, both from the book of Lugosi and Cesa-Bianchi [20]). For completeness, their pseudocodes are shown in Figures 8–9. Note that with Exp3.P (lines 6–14) we use the loss matrix transformation described earlier, hence the loss matrix has zero entries for the second (non-revealing) action, while the entry for action 1 and outcome j is $\ell_{1,j}(v) - \ell_{2,j}(v)$. Here $\ell_{i,j}(v)$ stands for the loss of action i and outcome j in the game $G(v)$ that is stored at node v .

function PLAY(v)**Input:** v is a tree node

```

1: if NUMOFActions( $G(v)$ ) = 2 then
2:    $(p, h) \leftarrow \text{PLAYATLEAF}(v)$ 
3: else
4:    $(p, h) \leftarrow \text{PLAY}(\text{CHILD}(v, g(v)))$ 
5:    $\hat{\rho}(v) \leftarrow (1 - \frac{1}{t(v)})\hat{\rho}(v) + \frac{1}{t(v)} \frac{\mathbb{I}(h=2)}{p}$ 
6:   if  $g(v) = 2$  and  $\hat{\rho}(v) < \rho'_1(v)$  then
7:     RESET(CHILD( $v, 1$ ));  $g(v) \leftarrow 1$ 
8:   else if  $g(v) = 1$  and  $\hat{\rho}(v) > \rho'_2(v)$  then
9:     RESET(CHILD( $v, 2$ ));  $g(v) \leftarrow 2$ 
10:  end if
11:   $t(v) \leftarrow t(v) + 1$ 
12: end if
13: return  $(p, h)$ 

```

Figure 6: The recursive function PLAY

function RESET(v)**Input:** v is a tree node

```

1: if NUMOFActions( $G(v)$ ) = 2 then
2:    $w_i(v) \leftarrow 1/2, i \leftarrow 1, 2$ 
3: else
4:    $g(v) \leftarrow 1, \hat{\rho}(v) \leftarrow 0, t(v) \leftarrow 1$ 
5:   RESET(CHILD( $v, 1$ ))
6: end if

```

Figure 7: Function RESET

5.2. Proof of the upper bound

Theorem 10. Assume $G = (\mathbf{L}, \mathbf{H})$ satisfies the separation condition and the non-degeneracy condition and $\ell_{i,j} \leq 1$. Denote by \bar{R}_T the regret of Algorithm APPLETREE up to time step T . There exist constants c, p such that for any $0 < \delta < 1$ and $T \in \mathbb{N}$, for any outcome sequence J_1, \dots, J_T , the algorithm with input G, T, δ achieves $\Pr[\bar{R}_T \leq c\sqrt{T} \ln^p(2T/\delta)] \geq 1 - \delta$.

Throughout the proof we will analyze the algorithm's behavior at the root node. We will use time indices as follows. Let us define the filtration $\{\mathcal{F}_t = \sigma(I_1, \dots, I_t)\}_t$, where I_t is the action the algorithm plays at time step t . To any variable $x(v)$ used by the algorithm, we denote by $x_t(v)$ the value of $x(v)$ that is measurable with respect to \mathcal{F}_t , but not measurable with respect to \mathcal{F}_{t-1} . From now on we abbreviate $x_t(\text{root})$ by x_t . We start with two lemmas. The first lemma shows that the number of switches the algorithm makes is small.

Lemma 11. Let S be the number of times APPLETREE calls RESET at the root node. Then there exists a universal constant c^* such that $S \leq \frac{c^* \ln T}{\Delta}$, where $\Delta = \rho'_2 - \rho'_1$ with ρ'_1 and ρ'_2 given by (2).

Note that here we use the non-degeneracy condition to ensure that $\Delta > 0$.

Proof. Let s be the number of times the algorithm switches from G_2 to G_1 . Let $t_1 < \dots < t_s$ be the time steps when $\hat{\rho}_t$ becomes smaller than ρ'_1 . Similarly, let $t'_1 < \dots < t'_{s+\xi}$, ($\xi \in \{0, 1\}$) be the time steps when $\hat{\rho}_t$ becomes greater than ρ'_2 . Note that for all $1 \leq j < s$, $t'_j < t_j < t'_{j+1}$. Finally, for every $1 \leq j < s$, we define $t''_j = \min\{t \mid t'_j \leq t \leq t_j, (\forall t \leq \tau \leq t_j : \hat{\rho}_\tau \leq 1)\}$. In other words, t''_j is the time step when $\hat{\rho}_t$ drops below 1 and stays there until the next reset.

First we observe that if $t''_j \geq 2/\Delta$ then $\hat{\rho}_{t''_j} \geq (\rho'_1 + \rho'_2)/2$. Indeed, if $t''_j = t'_j$ then $\hat{\rho}_{t''_j} \geq \rho'_2$, on the other hand, if $t''_j \neq t'_j$ then $\hat{\rho}_{t''_j-1} > 1$ and, from the update rule we have

$$\hat{\rho}_{t''_j} = \left(1 - \frac{1}{t''_j}\right) \hat{\rho}_{t''_j-1} + \frac{1}{t''_j} \cdot \frac{\mathbb{I}(J_{t''_j} = 2)}{p_{t''_j}} \geq 1 - \frac{\Delta}{2} \geq \frac{\rho'_1 + \rho'_2}{2}.$$

function PLAYATLEAF(v)

Input: v is a tree node

```

1: if REVEALINGACTIONNUMBER( $G(v)$ ) = 2 then
    Full-information case
2:    $(p, h) \leftarrow \text{EWA}(v)$ 
3: else
    Partial-information case
4:    $p \leftarrow (1 - \gamma(v)) \frac{w_1(v)}{w_1(v) + w_2(v)} + \gamma(v)/2$ 
5:    $U \sim \mathcal{U}_{[0,1]}$   $\triangleright U$  is uniform in  $[0, 1]$ 
6:   if  $U < p$  then  $\triangleright$  Play revealing action
7:      $h \leftarrow \text{CHOOSE}(1)$   $\triangleright h \in \{1, 2\}$ 
8:      $L_1 \leftarrow (\ell_{1,h}(v) - \ell_{2,h}(v) + \beta(v))/p$ 
9:      $L_2 \leftarrow \beta(v)/(1 - p)$ 
10:     $w_1(v) \leftarrow w_1(v) \exp(-\eta(v)L_1)$ 
11:     $w_2(v) \leftarrow w_2(v) \exp(-\eta(v)L_2)$ 
12:   else
13:      $h \leftarrow \text{CHOOSE}(2)$   $\triangleright$  here  $h = *$ 
14:   end if
15: end if
16: return  $(p, h)$ 

```

Figure 8: Function PLAYATLEAF

function EWA(v)

Input: v is a tree node

```

1:  $p \leftarrow \frac{w_1(v)}{w_1(v) + w_2(v)}$ 
2:  $U \sim \mathcal{U}_{[0,1]}$   $\triangleright U$  is uniform in  $[0, 1]$ 
3: if  $U < p$  then
4:    $I \leftarrow 1$ 
5: else
6:    $I \leftarrow 2$ 
7: end if
8:  $h \leftarrow \text{CHOOSE}(I)$   $\triangleright h \in \{1, 2\}$ 
9:  $w_1(v) \leftarrow w_1(v) \exp(-\eta(v)\ell_{1,h}(v))$ 
10:  $w_2(v) \leftarrow w_2(v) \exp(-\eta(v)\ell_{2,h}(v))$ 
11: return  $(p, h)$ 

```

Figure 9: Function Ewa

The number of times the algorithm resets is at most $2s + 1$. Let j^* be the first index such that $t''_{j^*} \geq 2/\Delta$. For any $j^* \leq j \leq s$, $\hat{\rho}'_{t'_j} \geq (\rho'_1 + \rho'_2)/2$ and $\hat{\rho}_{t_j} \leq \rho'_1$. According to the update rule we have for any $t'_j < t \leq t_j$ that

$$\hat{\rho}_t = \left(1 - \frac{1}{t}\right) \hat{\rho}_{t-1} + \frac{1}{t} \cdot \frac{\mathbb{I}(J_t = 2)}{p_t} \geq \hat{\rho}_{t-1} - \frac{1}{t} \hat{\rho}_{t-1} \geq \hat{\rho}_{t-1} - \frac{1}{t}$$

and hence $\hat{\rho}_{t-1} - \hat{\rho}_t \leq \frac{1}{t}$. Summing this inequality for all $t'_j + 1 \leq t \leq t_j$ such that $j \geq j^*$ we get

$$\begin{aligned} \frac{\Delta}{2} &= \frac{\rho'_1 + \rho'_2}{2} - \rho'_1 \leq \hat{\rho}'_{t'_j} - \hat{\rho}_{t_j} \\ &\leq \sum_{t=t'_j+1}^{t_j} \frac{1}{t} = O\left(\ln \frac{t_j}{t'_j}\right). \end{aligned}$$

Thus, there exists $c > 0$ such that for all $j^* \leq j \leq s$

$$\frac{1}{c} \Delta \leq \ln \frac{t_j}{t'_j} \leq \ln \frac{t_j}{t_{j-1}}. \quad (3)$$

Adding (3) for $j^* < j \leq s$ we get $(s - j^*) \frac{1}{c} \Delta \leq \ln \frac{t_s}{2/\Delta} \leq \ln T$. We conclude the proof with observing that $j^* \leq 2/\Delta$. \square

The next lemma shows that the estimate of the relative frequency of outcome 2 is not far away from its true value.

Lemma 12. For any $0 < \delta < 1$, with probability at least $1 - \delta$, for all $t \geq 8 \sqrt{T} \ln(2T/\delta)/(3\Delta^2)$, $|\hat{\rho}_t - \rho_t| \leq \Delta$.

The proof of the lemma employs Bernstein's inequality for martingales.

Bernstein's inequality for martingales. [20, Lemma A.8] Let X_1, X_2, \dots, X_n be a bounded martingale difference sequence with respect to a filtration $\{\mathcal{F}_i\}_{i=0}^n$ and with $|X_i| \leq K$. Let

$$S_i = \sum_{j=1}^i X_j$$

be the associated martingale. Denote the sum of conditional variances by

$$\Sigma_n^2 = \sum_{i=1}^n \mathbf{E}[X_i^2 \mid \mathcal{F}_{i-1}] .$$

Then, for all constants $\epsilon, \nu > 0$,

$$\Pr \left[\max_{i \in \underline{n}} S_i > \epsilon \text{ and } \Sigma_n^2 \leq \nu \right] \leq \exp \left(-\frac{\epsilon^2}{2(\nu + K\epsilon/3)} \right) .$$

Proof of Lemma 12. For $1 \leq t \leq T$, let p_t be the conditional probability of playing a revealing action at time step t , given the history $\mathcal{H}_{1:t-1}$. Recall that, due to the construction of the algorithm, $p_t \geq 1/\sqrt{T}$.

If we write $\hat{\rho}_t$ in its explicit form $\hat{\rho}_t = \frac{1}{t} \sum_{s=1}^t \frac{\mathbb{I}(H_s=2)}{p_s}$ we can observe that $\mathbf{E}[\hat{\rho}_t \mid \mathcal{H}_{1:t-1}] = \rho_t$, that is, $\hat{\rho}_t$ is an unbiased estimate of the relative frequency. Let us define random variables $X_s := \frac{\mathbb{I}(H_s=2)}{p_s} - \mathbb{I}(J_s = 2)$. Since p_s is determined by the history, $\{X_s\}_s$ is a martingale difference sequence. Also, from $p_s \geq 1/\sqrt{T}$ we know that $\mathbf{Var}(X_s \mid \mathcal{H}_{1:t-1}) \leq \sqrt{T}$. Hence, we can use Bernstein's inequality for martingales with $\epsilon = \Delta t$, $\nu = t\sqrt{T}$, $K = \sqrt{T}$:

$$\begin{aligned} \Pr [|\hat{\rho}_t - \rho_t| > \Delta] &= \Pr \left[\left| \sum_{s=1}^t X_s \right| > t\Delta \right] \\ &\leq 2 \exp \left(-\frac{\Delta^2 t^2 / 2}{t\sqrt{T} + \Delta t\sqrt{T}/3} \right) \\ &\leq 2 \exp \left(-\frac{3\Delta^2 t}{8\sqrt{T}} \right) . \end{aligned}$$

We have that if $t \geq 8\sqrt{T} \ln(2T/\delta)/(3\Delta^2)$ then

$$\Pr [|\hat{\rho}_t - \rho_t| > \Delta] \leq \delta/T .$$

We get the bound for all $t \in [8\sqrt{T} \ln(2T/\delta)/(3\Delta^2), T]$ using the union bound. □

Proof of Theorem 10. To prove that the algorithm achieves the desired regret bound we use induction on the depth of the tree, d . If $d = 1$, APPLE TREE plays either EWA or Exp3.P. EWA is known to satisfy Theorem 10, and, as we discussed earlier, Exp3.P achieves $O(\sqrt{T} \ln T/\delta)$ regret as well. As the induction hypothesis we assume that Theorem 10 is true for any T and any game such that the tree built by the algorithm has depth $d' < d$.

Let $Q_1 = \{1, \dots, \lceil K/2 \rceil\}$, $Q_2 = \{\lceil K/2 \rceil, \dots, K\}$ be the sets of actions associated with the subgames in the root. (Recall that the actions are ordered with respect to $\ell_{\cdot,1}$.) Furthermore, let us define the following

values: Let $T_0^0 = 1$, let T_i^0 be the first time step t after T_{i-1}^0 such that $g_t \neq g_{t-1}$. In other words, T_i^0 are the time steps when the algorithm switches between the subgames. Finally, let $T_i = \min(T_i^0, T + 1)$. From Lemma 11 we know that $T_{S_{\max}+1} = T + 1$, where $S_{\max} = \frac{c^* \ln T}{\Delta}$. It is easy to see that T_i are stopping times for any $i \geq 1$.

Without loss of generality, from now on we will assume that the optimal action $i^* \in Q_1$. If $i^* = \lceil K/2 \rceil$ then, since it is contained in both subgames, the bound trivially follows from the induction hypothesis and Lemma 11. In the rest of the proof we assume $i^* < K/2$.

Let $S = \max\{i \geq 1 \mid T_i^0 \leq T\}$ be the number of switches, $c = \frac{8}{3\Delta^2}$, and \mathcal{B} be the event that for all $t \geq c\sqrt{T} \ln(4T/\delta)$, $|\hat{\rho}_t - \rho_t| \leq \Delta$. We know from Lemma 12 that $\Pr[\mathcal{B}] \geq 1 - \delta/2$. On \mathcal{B} we have that $|\hat{\rho}_T - \rho_T| \leq \Delta$, and thus, using that $i^* < K/2$, $\rho_T \leq \rho_1^*$. This implies that in the last phase the algorithm plays on G_1 . It is also easy to see that before the last switch, at time step $T_S - 1$, $\hat{\rho}$ is between ρ_1^* and ρ_2^* , if T_S is large enough. Thus, up to time step $T_S - 1$, the optimal action is $\lceil K/2 \rceil$, the one that is shared by the two subgames. This implies that $\sum_{t=1}^{T_S-1} \ell_{i^*, J_t} - \ell_{\lceil K/2 \rceil, J_t} \geq 0$. On the other hand, if $T_S \leq c\sqrt{T} \ln(4T/\delta)$ then

$$\sum_{t=1}^{T_S-1} \ell_{i^*, J_t} - \ell_{\lceil K/2 \rceil, J_t} \geq -c\sqrt{T} \ln(4T/\delta).$$

Thus, we have

$$\begin{aligned} \widehat{R}_T &= \sum_{t=1}^T \ell_{I_t, J_t} - \ell_{i^*, J_t} \\ &= \sum_{t=1}^{T_S-1} (\ell_{I_t, J_t} - \ell_{i^*, J_t}) + \sum_{t=T_S}^T (\ell_{I_t, J_t} - \ell_{i^*, J_t}) \\ &\leq \mathbb{I}(\mathcal{B}) \left(\sum_{t=1}^{T_S-1} (\ell_{I_t, J_t} - \ell_{\lceil K/2 \rceil, J_t}) + \sum_{t=T_S}^T (\ell_{I_t, J_t} - \ell_{i^*, J_t}) \right) \\ &\quad + \underbrace{c\sqrt{T} \ln(4T/\delta) + (\mathbb{I}(\mathcal{B}^c))T}_D \\ &\leq D + \mathbb{I}(\mathcal{B}) \sum_{r=1}^{S_{\max}} \max_{i \in Q_{\pi(r)}} \sum_{t=T_{r-1}}^{T_r-1} (\ell_{I_t, J_t} - \ell_{i, J_t}) \\ &= D + \mathbb{I}(\mathcal{B}) \sum_{r=1}^{S_{\max}} \max_{i \in Q_{\pi(r)}} \sum_{m=1}^T \mathbb{I}(T_r - T_{r-1} = m) \sum_{t=T_{r-1}}^{T_{r-1}+m-1} (\ell_{I_t, J_t} - \ell_{i, J_t}), \end{aligned}$$

where $\pi(r)$ is 1 if r is odd and 2 if r is even. Note that for the last line of the above inequality chain to be well defined, we need outcome sequences of length at most $2T$. It does us no harm to assume that for all $T < t \leq 2T$, say, $J_t = 1$.

Recall that the strategies that play in the subgames are reset after the switches. Hence, the sum $\widehat{R}_m^{(r)} = \sum_{t=T_{r-1}}^{T_{r-1}+m-1} (\ell_{I_t, J_t} - \ell_{i, J_t})$ is the regret of the algorithm if it is used in the subgame $G_{\pi(r)}$ for $m \leq T$ steps. Then, exploiting that T_r are stopping times, we can use the induction hypothesis to bound $\widehat{R}_m^{(r)}$. In particular, let \mathcal{C} be the event that for all $m \leq T$ the sum is less than $c\sqrt{T} \ln^p(2T^2/\delta)$. Since the root node calls its children

with confidence parameter $\delta/(2T)$, we have that $\Pr[C^c] \leq \delta/2$. In summary,

$$\begin{aligned}\widehat{R}_T &\leq D + \mathbb{I}(C^c)T + \mathbb{I}(\mathcal{B})\mathbb{I}(C)S_{\max}c\sqrt{T}\ln^p 2T^2/\delta \\ &\leq \mathbb{I}(\mathcal{B}^c \cup C^c)T + c\sqrt{T}\ln(4T/\delta) + \mathbb{I}(\mathcal{B})\mathbb{I}(C)\frac{c^*\ln T}{\Delta}c\sqrt{T}\ln^p 2T^2/\delta.\end{aligned}$$

Thus, on $\mathcal{B} \cap C$, $\widehat{R}_T \leq \frac{2^p c c^*}{\Delta} \sqrt{T} \ln^{p+1}(2T/\delta)$, which, together with $\Pr[\mathcal{B}^c \cup C^c] \leq \delta$ concludes the proof. \square

Remark The above theorem proves a high probability bound on the regret. We can get a bound on the expected regret if we set δ to $1/\sqrt{T}$. Also note that the bound given by the induction grows in the number of non-dominated actions as $O(K^{\log_2 K})$.

6. Lower Bound for Non-Trivial Games

In the following sections, $\|\cdot\|_1$ and $\|\cdot\|$ denote the L_1 - and L_2 -norm of a vector in a Euclidean space, respectively.

In this section, we show that non-trivial games have minimax regret at least $\Omega(\sqrt{T})$. We state and prove this result for *all* finite games, in contrast to earlier related lower bounds which apply to specific losses (see Cesa-Bianchi and Lugosi [20, Theorems 3.7, 6.3, 6.4, 6.11] for full-information, label efficient, and bandit games).

Theorem 13 (Lower bound for non-trivial games). *If $G = (\mathbf{L}, \mathbf{H})$ is a finite non-trivial ($K \geq 2$) partial-monitoring game then there exists a constant $c > 0$ such that for any $T \geq 1$ the minimax expected regret $R_T(G) \geq c\sqrt{T}$.*

The proof presented below works for stochastic nature, as well. There is a far simpler proof in the Appendix, however, that one applies only for adversarial nature.

Recall that $\Delta_M \subset \mathbb{R}^M$ is the $(M-1)$ -dimensional probability simplex.

For the proof, we start with a geometrical lemma, which ensures the existence of a pair i_1, i_2 of non-dominated actions that are “neighbors” in the sense that for any small enough $\epsilon > 0$, there exists a pair of “ ϵ -close” outcome distributions $p + \epsilon v$ and $p - \epsilon v$ such that i_1 is uniquely optimal under the first distribution, and i_2 is uniquely optimal under the second distribution overtaking each non-optimal action by at least $\Omega(\epsilon)$ in both cases.

Lemma 14 (ϵ -close distributions). *Let $G = (\mathbf{L}, \mathbf{H})$ be any finite non-trivial game with N non-duplicate actions and $M \geq 2$ outcomes. Then there exist two non-dominated actions $i_1, i_2 \in \underline{N}$, $p \in \Delta_M$, $v \in \mathbb{R}^M \setminus \{0\}$, and $c, \alpha > 0$ satisfying the following properties:*

- (a) $\ell_{i_1} \neq \ell_{i_2}$.
- (b) $\langle \ell_{i_1}, p \rangle = \langle \ell_{i_2}, p \rangle \leq \langle \ell_i, p \rangle$ for all $i \in \underline{N}$ and the coordinates of p are positive.
- (c) Coordinates of v satisfy $\sum_{j=1}^M v(j) = 0$.

For any $\epsilon \in (0, \alpha)$,

- (d) $p_1 = p + \epsilon v \in \Delta_M$ and $p_2 = p - \epsilon v \in \Delta_M$,
- (e) for any $i \in \underline{N}$, $i \neq i_1$, we have $\langle \ell_i - \ell_{i_1}, p_1 \rangle \geq c\epsilon$,

(f) for any $i \in \underline{N}$, $i \neq i_2$, we have $\langle \ell_i - \ell_{i_2}, p_2 \rangle \geq c\epsilon$.

Proof of Lemma 14. For any action $i \in \underline{N}$, consider the cell

$$C_i = \{p \in \Delta_M : \forall i' \in \underline{N}, \langle \ell_i, p \rangle \leq \langle \ell_{i'}, p \rangle\}$$

in the probability simplex Δ_M . The cell C_i corresponds to the set of outcome distributions under which action i is optimal. Each cell is the intersection of some closed half-spaces and Δ_M , and thus it is a compact convex polytope of dimension at most $M - 1$. Note that

$$\bigcup_{i=1}^N C_i = \Delta_M. \quad (4)$$

For $C \subseteq \Delta_M$, denote $\text{int } C$ its interior in the topology induced by the hyperplane $\{x \in \mathbb{R}^M : \langle (1, \dots, 1), x \rangle = 1\}$ and $\text{rint } C$ its relative interior⁶. Let λ be the $(M - 1)$ -dimensional Lebesgue-measure. It is easy to see that for any pair of cells $C_i, C_{i'}$, $C_{i'} \cap \text{int } C_i = \emptyset$, that is, $\lambda(C_i \cap C_{i'}) = 0$, and so

$$\text{int } C_i \subseteq C_i \setminus \bigcup_{i' \neq i} C_{i'}. \quad (5)$$

Hence the cells form a cell-decomposition of the simplex. Any two cells C_i and $C_{i'}$ are separated by the hyperplane $f_{i,i'} = \{x \in \mathbb{R}^M : \langle \ell_i, x \rangle = \langle \ell_{i'}, x \rangle\}$. Note that $C_i \cap C_{i'} \subset f_{i,i'}$. The cells are characterized by the following lemma (which itself holds also with duplicate actions):

Lemma 15. *Action i is dominated $\Leftrightarrow C_i \subseteq \bigcup_{i': \ell_{i'} \neq \ell_i} C_{i'} \Leftrightarrow \text{int } C_i = \emptyset \Leftrightarrow \lambda(C_i) = 0$, that is, C_i is $(M - 1)$ -dimensional (has positive λ -measure) if and only if there is $p \in C_i \setminus \bigcup_{i': \ell_{i'} \neq \ell_i} C_{i'}$. Hence there is three kind of “cells”:*

1. $C_i = \emptyset$ (action i is never optimal),
2. $C_i \neq \emptyset$ has dimension less than $M - 1$, $\text{int } C_i = \emptyset$, $\lambda(C_i) = 0$, $C_i \subseteq \bigcup_{i': \ell_{i'} \neq \ell_i} C_{i'}$ (action i is degenerate),
3. action i is non-dominated, C_i is $(M - 1)$ -dimensional, $\text{rint } C_i = \text{int } C_i \neq \emptyset$, $\lambda(C_i) > 0$, there is $p \in C_i \setminus \bigcup_{i': \ell_{i'} \neq \ell_i} C_{i'}$.

Moreover $\bigcup_{i \notin \mathcal{D}} C_i = \Delta_M$ for the set \mathcal{D} of dominated actions.

The proof is in the Appendix.

The non-triviality of the game ($K \geq 2$) means that there are at least two non-dominated actions of type 3 above. In the cell decomposition, due to Lemma 15, there must exist two such $(M - 1)$ -dimensional cells C_{i_1} and C_{i_2} corresponding to two non-dominated actions i_1, i_2 , such that their intersection $C_{i_1} \cap C_{i_2}$ is an $(M - 2)$ -dimensional polytope. Clearly, $\ell_{i_1} \neq \ell_{i_2}$, since otherwise the cells would coincide; thus part (a) is satisfied.

Moreover, $\text{rint}(C_{i_1} \cap C_{i_2}) \subseteq \text{rint } \Delta_M$ since otherwise $\lambda(C_{i_1})$ or $\lambda(C_{i_2})$ would be zero. We can choose any $p \in \text{rint}(C_{i_1} \cap C_{i_2})$. This choice of p guarantees that $p \in f_{i_1, i_2}$, $\langle \ell_{i_1}, p \rangle = \langle \ell_{i_2}, p \rangle$, $p \in \text{rint } \Delta_M$, and part (b) is satisfied. Since $C_{i_1} \cap C_{i_2}$ is $(M - 2)$ -dimensional, it also implies that there exists $\delta > 0$ such that the δ -neighborhood $\{q \in \mathbb{R}^M : \|p - q\| < \delta\}$ of p is contained in $\text{rint}(C_{i_1} \cup C_{i_2})$.

Since $p \in f_{i_1, i_2}$ therefore the hyperplane of vectors satisfying (c) does not coincide with f_{i_1, i_2} implying that we can choose $v \in \mathbb{R}^M \setminus \{0\}$ satisfying part (c), $\|v\| < \delta$, and $v \notin f_{i_1, i_2}$. We can assume

$$\langle \ell_{i_2} - \ell_{i_1}, v \rangle > 0 \quad (6)$$

⁶Relative interior of $C \subseteq \mathbb{R}^M$ is its interior in the topology induced by the smallest affine space containing it.

(otherwise we choose $-v$). Since $p \pm v$ lie in the δ -neighborhood of p , they lie in $\text{rint}(C_{i_1} \cup C_{i_2})$. In particular, since $\langle \ell_{i_1}, p + v \rangle < \langle \ell_{i_2}, p + v \rangle$ and $\langle \ell_{i_2}, p - v \rangle < \langle \ell_{i_1}, p - v \rangle$, $p + v \in \text{rint } C_{i_1}$ and $p - v \in \text{rint } C_{i_2}$. Let

$$p_1 = p + \epsilon v \quad \text{and} \quad p_2 = p - \epsilon v. \quad (7)$$

The convexity of C_{i_1} and C_{i_2} implies that for any $\epsilon \in (0, 1]$, $p_1 \in \text{rint } C_{i_1}$ and $p_2 \in \text{rint } C_{i_2}$. This, in particular, ensures that $p_1, p_2 \in \Delta_M$ and part (d) holds.

To prove (e) define $\mathcal{I} = \{i \in \underline{N} : \ell_i \text{ is collinear with } \ell_{i_1} \text{ and } \ell_{i_2}\}$. We consider two cases: As the first case fix action $i \in \mathcal{I} \setminus \{i_1\}$, that is, ℓ_i is an affine combination $\ell_i = a_i \ell_{i_1} + b_i \ell_{i_2}$ for some $a_i + b_i = 1$. Since i_1 and i_2 are non-dominated, this must be a convex combination with $a_i, b_i \geq 0$. There is no duplicate action, thus $\ell_i \neq \ell_{i_1}$ implying $b_i \neq 0$. Hence $b_i > 0$, and from (7) for any $\epsilon \geq 0$

$$\langle \ell_i - \ell_{i_1}, p_1 \rangle = \langle b_i \ell_{i_2} - b_i \ell_{i_1}, p + \epsilon v \rangle = \epsilon b_i \langle \ell_{i_2} - \ell_{i_1}, v \rangle \geq c \epsilon$$

provided that $0 < c \leq \min_{i \in \mathcal{I} \setminus \{i_1\}} b_i \langle \ell_{i_2} - \ell_{i_1}, v \rangle = c'$. From (6) we know that $b_i \langle \ell_{i_2} - \ell_{i_1}, v \rangle$ and so c' are positive.

As the second case suppose $i \notin \mathcal{I}$. Then, the hyperplane $f_{i_1, i}$ does not coincide with f_{i_1, i_2} . Since $p \in \text{rint}(C_{i_1} \cap C_{i_2})$, $p \in f_{i_1, i}$ would contradict to $f_{i_1, i} \cap \text{rint } C_{i_1} = \emptyset$ implied by (5). Thus $p \in C_{i_1} \setminus f_{i_1, i}$ and therefore $\langle \ell_{i_1}, p \rangle < \langle \ell_i, p \rangle$. This means that if we choose $0 < c \leq \min(c', \frac{1}{2} \min_{i \notin \mathcal{I}} \langle \ell_i - \ell_{i_1}, p \rangle)$ (that is positive and depends only on \mathbf{L} and not on T) then for $\epsilon < \alpha = \min(1, c / \max_{i \notin \mathcal{I}} |\langle \ell_i - \ell_{i_1}, v \rangle|)$, from (7) we have again

$$\langle \ell_i - \ell_{i_1}, p_1 \rangle \geq 2c + \epsilon \langle \ell_i - \ell_{i_1}, v \rangle > c > c \epsilon.$$

Part (f) is proved analogously to part (e), and by adjusting α and c if necessary. \square

We now continue with a technical lemma, which quantifies an upper bound on the Kullback-Leibler (KL) divergence (or relative entropy) between the two distributions from the previous lemma. Recall that the KL divergence between two probability distributions $p, q \in \Delta_M$ is defined as

$$D(p \parallel q) = \sum_{j=1}^M p_j \ln \left(\frac{p_j}{q_j} \right).$$

Lemma 16 (KL divergence of ϵ -close distributions). *Let $p \in \Delta_M$ be a probability vector. For any vector $\epsilon \in \mathbb{R}^M$ such that both $p - \epsilon$ and $p + \epsilon$ lie in Δ_M and $|\epsilon(j)| \leq p(j)/2$ for all $j \in \underline{M}$, the KL divergence of $p - \epsilon$ and $p + \epsilon$ satisfies*

$$D(p - \epsilon \parallel p + \epsilon) \leq c \|\epsilon\|^2$$

for some constant c depending only on p .

Proof of Lemma 16. Since $p, p + \epsilon$, and $p - \epsilon$ are all probability vectors, notice that the coordinates of ϵ have to sum up to zero. Also if a coordinate of p is zero then the corresponding coordinate of ϵ has to be zero as well. As zero coordinates do not modify the KL divergence, we can assume without loss of generality that all coordinates of p are positive. By definition,

$$D(p - \epsilon \parallel p + \epsilon) = \sum_{j=1}^M (p(j) - \epsilon(j)) \ln \left(\frac{p(j) - \epsilon(j)}{p(j) + \epsilon(j)} \right).$$

We write the logarithmic factor as

$$\ln \left(\frac{p(j) - \epsilon(j)}{p(j) + \epsilon(j)} \right) = \ln \left(1 - \frac{\epsilon(j)}{p(j)} \right) - \ln \left(1 + \frac{\epsilon(j)}{p(j)} \right).$$

We use the second order Taylor expansion $\ln(1 \pm x) = \pm x - x^2/2 + O(|x|^3)$ around 0 to get that $\ln(1 - x) - \ln(1 + x) = -2x + r(x)$, where $r(x)$ is a remainder upper bounded for all $|x| \leq 1/2$ as $|r(x)| \leq c'|x|^3$ with some universal constant $c' > 0$.⁷ Substituting

$$\begin{aligned} D(p - \varepsilon \parallel p + \varepsilon) &= \sum_{j=1}^M (p(j) - \varepsilon(j)) \left[-2 \frac{\varepsilon(j)}{p(j)} + r\left(\frac{\varepsilon(j)}{p(j)}\right) \right] \\ &= -2 \sum_{j=1}^M \varepsilon(j) + 2 \sum_{j=1}^M \frac{\varepsilon^2(j)}{p(j)} + \sum_{j=1}^M (p(j) - \varepsilon(j)) \cdot r\left(\frac{\varepsilon(j)}{p(j)}\right). \end{aligned}$$

Here the first term is 0. Letting $\underline{p} = \min_{j \in M} p(j)$, the second term is bounded by $2 \sum_{j=1}^M \varepsilon^2(j)/\underline{p} = (2/\underline{p})\|\varepsilon\|^2$, and the third term is bounded by

$$\begin{aligned} \sum_{j=1}^M (p(j) - \varepsilon(j)) \left| r\left(\frac{\varepsilon(j)}{p(j)}\right) \right| &\leq c' \sum_{j=1}^M (p(j) - \varepsilon(j)) \frac{|\varepsilon(j)|^3}{p^3(j)} = c' \sum_{j=1}^M \left(\frac{|\varepsilon(j)|}{p(j)} - \frac{\varepsilon(j)|\varepsilon(j)|}{p^2(j)} \right) \frac{\varepsilon^2(j)}{p(j)} \\ &\leq c' \sum_{j=1}^M \left(\frac{|\varepsilon(j)|}{p(j)} + \frac{|\varepsilon(j)|^2}{p^2(j)} \right) \frac{\varepsilon^2(j)}{p(j)} \\ &\leq c' \sum_{j=1}^M \left(\frac{1}{2} + \frac{1}{4} \right) \frac{\varepsilon^2(j)}{\underline{p}} = \frac{3c'}{4\underline{p}} \|\varepsilon\|^2. \end{aligned}$$

Hence, $D(p - \varepsilon \parallel p + \varepsilon) \leq \frac{8+3c'}{4\underline{p}} \|\varepsilon\|^2 = c \|\varepsilon\|^2$ for $c = \frac{8+3c'}{4\underline{p}}$. \square

Proof of Theorem 13. The proof is similar as in Auer et al. [5]. When $M = 1$, G is always trivial, thus we assume that $M \geq 2$. Without loss of generality we may assume that all the actions are all-revealing. Then, as in Section 3 for $M=2$, we can also assume that there are no duplicate actions, thus for any two actions i and i' , $\ell_i \neq \ell_{i'}$.

Lemma 14 implies that there exist two actions i_1, i_2 , $p \in \Delta_M$, $v \in \mathbb{R}^M$, and $c_1, \alpha > 0$ satisfying conditions (a)–(f). To avoid cumbersome indexing, by renaming the actions we can achieve that $i_1 = 1$ and $i_2 = 2$. Let $p_1 = p + \epsilon v$ and $p_2 = p - \epsilon v$ for some $\epsilon \in (0, \alpha)$. We determine the precise value of ϵ later. By Lemma 14 (d), $p_1, p_2 \in \Delta_M$.

Fix any randomized learning algorithm A and time horizon T . We use randomization replacing the outcomes by a sequence J_1, J_2, \dots, J_T of random variables i.i.d. according to p_k , $k \in \{1, 2\}$, and independently of the internal randomization of A . Let

$$N_i^{(k)} = N_i^{(k)}(A, T) = \sum_{t=1}^T \Pr_k[I_t = i] \in [0, T] \quad (8)$$

be the expected number of times action i is chosen by A under p_k up to time step T . With subindex k , \Pr_k and \mathbf{E}_k denote probability and expectation given outcome model $k \in \{1, 2\}$, respectively.

Lemma 17. *For any partial-monitoring game with N actions and M outcomes, algorithm A and outcome distribution $p_k \in \Delta_M$ such that action k is optimal under p_k , we have*

$$\bar{R}_T(A, G) \geq \sum_{\substack{i \in N \\ i \neq k}} N_i^{(k)} \langle \ell_i - \ell_k, p_k \rangle, \quad k = 1, 2. \quad (9)$$

⁷In fact, one can take $c' = 8 \ln(3/e) \approx 0.79$.

The proof is in the Appendix.

Parts (e) and (f) of Lemma 14 imply that $\langle \ell_k, p_k \rangle \leq \langle \ell_i, p_k \rangle$ for $k \in \{1, 2\}$ and any $i \in \underline{N}$, hence $\bar{R}_T(A, G)$ can be bounded in terms of $N_i^{(k)}$ using Lemma 17. They also imply that for any $i \in \underline{N}$ if $\ell_i \neq \ell_k$ then $\langle \ell_i - \ell_k, p_k \rangle \geq c_1 \epsilon$. Therefore, we can continue lower bounding (9) as

$$\sum_{\substack{i \in \underline{N} \\ i \neq k}} N_i^{(k)} \langle \ell_i - \ell_k, p_k \rangle \geq \sum_{\substack{i \in \underline{N} \\ i \neq k}} N_i^{(k)} c_1 \epsilon = c_1 (T - N_k^{(k)}) \epsilon. \quad (10)$$

Collecting (9) and (10), we see that the worst-case regret of A is lower bounded by

$$\bar{R}_T(A, G) \geq c_1 (T - N_k^{(k)}) \epsilon \quad (11)$$

for $k \in \{1, 2\}$. Averaging (11) over $k \in \{1, 2\}$ we get

$$\bar{R}_T(A, G) \geq c_1 (2T - N_1^{(1)} - N_2^{(2)}) \epsilon / 2. \quad (12)$$

We now focus on lower bounding $2T - N_1^{(1)} - N_2^{(2)}$. We start by showing that $N_2^{(2)}$ is close to $N_2^{(1)}$. The following lemma, which is the key lemma of both lower bound proofs, carries that out formally and states that the expected number of times an action is played by A does not change too much when we change the model, if the outcome distributions p_1 and p_2 are “close” in KL-divergence:

Lemma 18. *For any partial-monitoring game with N actions and M outcomes, algorithm A , pair of outcome distributions $p_1, p_2 \in \Delta_M$ and action i , we have*

$$N_i^{(2)} - N_i^{(1)} \leq T \sqrt{D(p_2 \parallel p_1) N_{\text{rev}}^{(2)} / 2} \quad \text{and} \quad N_i^{(1)} - N_i^{(2)} \leq T \sqrt{D(p_1 \parallel p_2) N_{\text{rev}}^{(1)} / 2},$$

where $N_{\text{rev}}^{(k)} = \sum_{t=1}^T \Pr_k[I_t \in \mathcal{R}] = \sum_{i \in \mathcal{R}} N_i^{(k)}$ under model p_k , $k = 1, 2$ with \mathcal{R} being the set of revealing actions.⁸

The proof is in the Appendix.

We use Lemma 18 for $i = 2$ and that $N_{\text{rev}}^{(2)} \leq T$ to bound the difference $N_2^{(2)} - N_2^{(1)}$ as

$$N_2^{(2)} - N_2^{(1)} \leq T \sqrt{D(p_2 \parallel p_1) T / 2} = T^{3/2} \sqrt{D(p_2 \parallel p_1) / 2}. \quad (13)$$

We upper bound $D(p_2 \parallel p_1)$ using Lemma 16 with $\epsilon = \epsilon v$. The lemma implies that $D(p_2 \parallel p_1) \leq c_2 \epsilon^2$ for $\epsilon < \epsilon_0$ with some $\epsilon_0, c_2 > 0$ which depend only on v and p . Putting this together with (13) we get

$$N_2^{(2)} < N_2^{(1)} + c_3 \epsilon T^{3/2}$$

where $c_3 = \sqrt{c_2 / 2}$. Together with $N_1^{(1)} + N_2^{(1)} \leq T$ we get

$$2T - N_1^{(1)} - N_2^{(2)} > 2T - N_1^{(1)} - N_2^{(1)} - c_3 \epsilon T^{3/2} \geq T - c_3 \epsilon T^{3/2}.$$

Substituting into (12) and choosing $\epsilon = 1/(2c_3 T^{1/2})$ gives the desired lower bound

$$\bar{R}_T(A, G) > \frac{c_1}{8c_3} \sqrt{T}$$

⁸It seems from the proof that $N_{\text{rev}}^{(k)}$ could be slightly sharpened to $N_{\text{rev}}^{(k, T-1)} = \sum_{t=1}^{T-1} \Pr_k[I_t \in \mathcal{R}]$.

provided that our choice of ϵ ensures that $\epsilon < \min(\alpha, \epsilon_0) =: \epsilon_1$ that depends only on \mathbf{L} . This condition is satisfied for all $T > T_0 = 1/(2c_3\epsilon_1)^2$. Since c_1 , c_3 , and ϵ_1 depend only on \mathbf{L} , for such T , $R_T(G) \geq \frac{c_1}{8c_3} \sqrt{T}$.

The non-triviality of the game implies that Lemma 3 d) does not hold, so neither does b), that is, $R_T(G) > 0$ for $T \geq 1$. Thus choosing

$$c = \min\left(\min_{1 \leq T \leq T_0} \frac{R_T(G)}{\sqrt{T}}, \frac{c_1}{8c_3}\right),$$

$c > 0$ and for any T , $R_T(G) \geq c \sqrt{T}$. □

Remark Theorem 13 also holds if $M = \infty$. Namely, since the proof of c) \Rightarrow d) of Lemma 3 remains obviously valid, the non-triviality of the game ($K \geq 2$) excludes that c) holds, and thus for each $i \in \underline{N}$ there is $j_i \in \{1, 2, \dots\}$ such that ℓ_{i,j_i} is not minimal in the j_i^{th} column of \mathbf{L} . Then take the minor of \mathbf{L} consisting of its (at most N) columns corresponding to $O = \{j_1, \dots, j_N\}$. For the corresponding finite game G_O (that does not depend on A), Lemma 3 c) still does not hold, thus nor d) does, and G_O is also non-trivial. Hence Theorem 13 implies that⁹

$$R_T(G) = \inf_A \sup_{j_1:T \in \{1,2,\dots\}^T} R_T(A, G) \geq \inf_A \sup_{j_1:T \in O^T} R_T(A, G) = R_T(G_O) = \Omega(\sqrt{T}).$$

7. Lower Bound for Hard Games

In this section, we present an $\Omega(T^{2/3})$ lower bound for the expected regret of any two-outcome game in the case when the separation condition does not hold.

Theorem 19 (Lower bound for hard games). *If $M = 2$ and $G = (\mathbf{L}, \mathbf{H})$ satisfies the non-degeneracy condition and the separation condition does **not** hold then there exists a constant $C > 0$ such that for any $T \geq 1$ the minimax expected regret $R_T(G) \geq CT^{2/3}$.*

Proof of Theorem 19. We follow the lower bound proof for the label efficient prediction from Cesa-Bianchi et al. [22] with a few changes. The most important change, as we will see, is the choice of the models we randomize over.

As the first step, the following lemma shows that non-revealing degenerate actions do not influence the minimax regret of a game.

Lemma 20. *Let G be a non-degenerate game with two outcomes. Let G' be the game we get by removing the degenerate non-revealing actions from G . Then $R_T(G) = R_T(G')$.*

The proof of this lemma can be found in the Appendix.

By the non-degeneracy condition and Lemma 20, we can assume without loss of generality that G does not have degenerate actions. We can also assume without loss of generality that actions 1 and 2 are the two consecutive non-dominated non-revealing actions. It follows by scaling and a reduction similar to the one we used in Section 5.1 that we can further assume $(\ell_{1,1}, \ell_{1,2}) = (0, \alpha)$, $(\ell_{2,1}, \ell_{2,2}) = (1 - \alpha, 0)$ with some $\alpha \in (0, 1)$. Using the non-degeneracy condition and that actions 1 and 2 are consecutive non-dominated actions, we get that for all $i \geq 3$, there exists some $\lambda_i \in \mathbb{R}$ depending only on \mathbf{L} such that

$$\begin{aligned} \ell_{i,1} &> \lambda_i \ell_{1,1} + (1 - \lambda_i) \ell_{2,1} = (1 - \lambda_i)(1 - \alpha), \\ \ell_{i,2} &> \lambda_i \ell_{1,2} + (1 - \lambda_i) \ell_{2,2} = \lambda_i \alpha. \end{aligned} \tag{14}$$

⁹The same reasoning can be used to show that we could assume without loss of generality $M \leq N$ in the proof of Theorem 13.

Let $\lambda_{\min} = \min_{i \geq 3} \lambda_i$, $\lambda_{\max} = \max_{i \geq 3} \lambda_i$, and $\lambda^* = \lambda_{\max} - \lambda_{\min}$.

We define two models for generating outcomes from $\{1, 2\}$. In model 1, the outcome distribution is $p_1(1) = \alpha + \epsilon$, $p_1(2) = 1 - p_1(1)$, whereas in model 2, $p_2(1) = \alpha - \epsilon$, $p_2(2) = 1 - p_2(1)$ with $0 < \epsilon \leq \min(\alpha, 1 - \alpha)/2$ to be chosen later. We use randomization replacing the outcomes by a sequence J_1, J_2, \dots, J_T of random variables i.i.d. according to p_k , $k \in \{1, 2\}$, and independently of the internal randomization of A . Let $N_i^{(k)}$ be the expected number of times action i is chosen by A under p_k up to time step T , as in (8). With subindex k , \Pr_k and \mathbf{E}_k denote probability and expectation given outcome model $k \in \{1, 2\}$, respectively. Finally, let $N_{\geq 3}^{(k)} = \sum_{i \geq 3} N_i^{(k)}$. Note that, if $\epsilon < \epsilon_0$ with some ϵ_0 depending only on \mathbf{L} then only actions 1 and 2 can be optimal for these models. Namely, action k is optimal under p_k , hence $\bar{R}_T(A, G)$ can be bounded in terms of $N_i^{(k)}$ using Lemma 17:

$$\bar{R}_T(A, G) \geq \sum_{\substack{i \in N \\ i \neq k}} N_i^{(k)} \langle \ell_i - \ell_k, p_k \rangle = \sum_{i=3}^N N_i^{(k)} \langle \ell_i - \ell_k, p_k \rangle + N_{3-k}^{(k)} \langle \ell_{3-k} - \ell_k, p_k \rangle \quad (15)$$

for $k = 1, 2$. Now, by (14), there exists $\tau > 0$ depending only on \mathbf{L} such that for all $i \geq 3$, $\ell_{i,1} \geq (1 - \lambda_i)(1 - \alpha) + \tau$ and $\ell_{i,2} \geq \alpha \lambda_i + \tau$. These bounds and simple algebra give that

$$\begin{aligned} \langle \ell_i - \ell_1, p_1 \rangle &= (\ell_{i,1} - \ell_{1,1})(\alpha + \epsilon) + (\ell_{i,2} - \ell_{1,2})(1 - \alpha - \epsilon) \\ &\geq ((1 - \lambda_i)(1 - \alpha) + \tau)(\alpha + \epsilon) + (\alpha \lambda_i + \tau - \alpha)(1 - \alpha - \epsilon) \\ &= (1 - \lambda_i)\epsilon + \tau \\ &\geq (1 - \lambda_{\max})\epsilon + \tau =: f_1 \end{aligned}$$

and

$$\langle \ell_2 - \ell_1, p_1 \rangle = (1 - \alpha)(\alpha + \epsilon) - \alpha(1 - \alpha - \epsilon) = \epsilon.$$

Analogously, we get

$$\langle \ell_i - \ell_2, p_2 \rangle \geq \lambda_{\min}\epsilon + \tau =: f_2 \quad \text{and} \quad \langle \ell_1 - \ell_2, p_2 \rangle = \epsilon.$$

Note that if $\epsilon < \tau / \max(|1 - \lambda_{\max}|, |\lambda_{\min}|)$ then both f_1 and f_2 are positive. Substituting these into (15) gives

$$\bar{R}_T(A, G) \geq f_k N_{\geq 3}^{(k)} + \epsilon N_{3-k}^{(k)}. \quad (16)$$

The following lemma is an application of Lemma 18 and 16:

Lemma 21. *There exists a constant $c > 0$ (depending on α only) such that*

$$N_2^{(1)} \geq N_2^{(2)} - cT\epsilon\sqrt{N_{\geq 3}^{(2)}} \quad \text{and} \quad N_1^{(2)} \geq N_1^{(1)} - cT\epsilon\sqrt{N_{\geq 3}^{(1)}}.$$

Proof. We only prove the first inequality, the other one is symmetric. Using Lemma 18 with $M = 2$, $i = 2$ and the fact that actions 1 and 2 are non-revealing, we have

$$N_2^{(2)} - N_2^{(1)} \leq T\sqrt{D(p_2 \parallel p_1)N_{\geq 3}^{(2)}/2}.$$

Lemma 16 with $M = 2$, $p = (\alpha, 1 - \alpha)^\top$, and $\epsilon = (\epsilon, -\epsilon)^\top$ gives $D(p_2 \parallel p_1) \leq \hat{c}\epsilon^2$, where \hat{c} depends only on α . Rearranging and substituting $c = \sqrt{\hat{c}/2}$ yields the first statement of the lemma. \square

Let $l = \arg \min_{k \in \{1,2\}} N_{\geq 3}^{(k)}$. Now, for $k \neq l$ we can lower bound the regret using Lemma 21 for (16):

$$\bar{R}_T(A, G) \geq f_k N_{\geq 3}^{(k)} + \epsilon \left(N_{3-k}^{(l)} - cT\epsilon \sqrt{N_{\geq 3}^{(l)}} \right) \geq f_k N_{\geq 3}^{(l)} + \epsilon \left(N_{3-k}^{(l)} - cT\epsilon \sqrt{N_{\geq 3}^{(l)}} \right), \quad (17)$$

as $f_k > 0$. For $k = l$ we do this subtracting $cT\epsilon^2 \sqrt{N_{\geq 3}^{(l)}} \geq 0$ from the right-hand side of (16) leading to the same lower bound, hence (17) holds for $k = 1, 2$. Finally, averaging (17) over $k \in \{1, 2\}$ we have the bound

$$\begin{aligned} \frac{f_1 + f_2}{2} N_{\geq 3}^{(l)} + \epsilon \left(\frac{N_2^{(l)} + N_1^{(l)}}{2} - cT\epsilon \sqrt{N_{\geq 3}^{(l)}} \right) &= \left(\frac{(1 - \lambda_{\max} + \lambda_{\min})\epsilon}{2} + \tau \right) N_{\geq 3}^{(l)} + \epsilon \left(\frac{T - N_{\geq 3}^{(l)}}{2} \right) - cT\epsilon^2 \sqrt{N_{\geq 3}^{(l)}} \\ &= \left(\tau - \frac{\lambda^* \epsilon}{2} \right) N_{\geq 3}^{(l)} + \frac{\epsilon T}{2} - cT\epsilon^2 \sqrt{N_{\geq 3}^{(l)}}. \end{aligned}$$

Choosing $\epsilon = c_2 T^{-1/3}$ ($\leq c_2$) with $c_2 > 0$ gives

$$\begin{aligned} \bar{R}_T(A, G) &\geq \left(\tau - \frac{\lambda^* c_2 T^{-1/3}}{2} \right) N_{\geq 3}^{(l)} + \frac{c_2 T^{2/3}}{2} - c c_2^2 T^{1/3} \sqrt{N_{\geq 3}^{(l)}} \\ &\geq \left(\tau - \frac{\lambda^* c_2}{2} \right) N_{\geq 3}^{(l)} + \frac{c_2 T^{2/3}}{2} - c c_2^2 T^{1/3} \sqrt{N_{\geq 3}^{(l)}} \\ &= \left(\left(\tau - \frac{\lambda^* c_2}{2} \right) x^2 + \frac{c_2}{2} - c c_2^2 x \right) T^{2/3} = q(x) T^{2/3}, \end{aligned}$$

where $x = T^{-1/3} \sqrt{N_{\geq 3}^{(l)}}$ and $q(x)$ can be written and lower bounded as

$$q(x) = \left(\tau - \frac{\lambda^* c_2}{2} \right) \left(x - \frac{c c_2^2}{2\tau - \lambda^* c_2} \right)^2 + \frac{c_2}{2} - \frac{c^2 c_2^4}{4\tau - 2\lambda^* c_2} \geq \frac{c_2}{2} \left(1 - \frac{c^2 c_2}{2\tau - \lambda^* c_2} \right)$$

independently of x whenever $\lambda^* c_2 < 2\tau$ and $c_2 \leq 1$. Now it is easy to see that if $c_2 = \min(\tau/(c^2 + \lambda^*), 1)$ then these hold, moreover, $q(x) \geq c_2/4 > 0$ giving the desired lower bound

$$\bar{R}_T(A, G) \geq \frac{c_2}{4} T^{2/3}$$

provided that our choice of ϵ ensures that $\epsilon < \min(\alpha/2, (1-\alpha)/2, \epsilon_0, \tau/|1-\lambda_{\max}|, \tau/|\lambda_{\min}|) =: \epsilon_1$ that depends only on \mathbf{L} . This condition is satisfied for all $T > T_0 = (c_2/\epsilon_1)^3$. Since c_2 and ϵ_1 depend only on \mathbf{L} , for such T , $R_T(G) \geq \frac{c_2}{4} T^{2/3}$.

If the separation condition does not hold then the game is clearly non-trivial which, using Lemma 3 b) and d) as in the proof of Theorem 13, implies that $R_T(G) > 0$ for $T \geq 1$. Thus choosing

$$C = \min \left(\min_{1 \leq T \leq T_0} \frac{R_T(G)}{T^{2/3}}, \frac{c_2}{4} \right),$$

$C > 0$ and for any T , $R_T(G) \geq C T^{2/3}$. □

8. Discussion

In this paper we classified non-degenerate partial-monitoring games with two outcomes based on their minimax regret. An immediate question is how the classification extends to degenerate games. Unfortunately, the degeneracy condition is needed in both the upper and lower bound proofs. We do not even know if all degenerate games fall into one of the four categories or there are some games with minimax regret of $\tilde{\Theta}(T^\alpha)$ for some $\alpha \in (1/2, 2/3)$. Nonetheless, we conjecture that, if the revealing degenerate actions are included in the chain of non-dominated actions, the classification theorem holds without any change.

The most important open question is whether our results generalize to games with more outcomes. A simple observation is that, given a finite partial-monitoring game, if we restrict the opponent's choices to any two outcomes, the resulting game's hardness serves as a lower bound on the minimax regret of the original game. This gives us a sufficient condition that a game has $\Omega(T^{2/3})$ minimax regret. We believe that the $\Omega(T^{2/3})$ lower bound can also be generalized to situations where two " ϵ -close" outcome distributions are not distinguishable by playing only their respective optimal actions. Generalizing the upper bound result seems more challenging. The algorithm APPLE TREE heavily exploits the two-dimensional structure of the losses and, as of yet, in general we do not know how to construct an algorithm that achieves $\tilde{O}(\sqrt{T})$ regret on partial-monitoring games with more than two outcomes.

It is also important to note that our upper bound result heavily exploits the assumption that the opponent is oblivious. Our results do not extend to games with non-oblivious opponents, to the best of our knowledge.

Appendix A.

Proof of Lemma 3. a) \Rightarrow b) is obvious.

b) \Rightarrow c) For any A ,

$$\begin{aligned} \bar{R}_T(A, G) &\geq \sup_{j \in \underline{M}, J_1 = \dots = J_T = j} \mathbf{E} \left[\sum_{t=1}^T \ell_{I_t, J_t} - \min_{i \in \underline{N}} \sum_{t=1}^T \ell_{i, J_t} \right] \\ &= \sup_{j \in \underline{M}} \mathbf{E} \left[\sum_{t=1}^T \ell_{I_t, j} - T \min_{i \in \underline{N}} \ell_{i, j} \right] \\ &\geq \sup_{j \in \underline{M}} \left(\mathbf{E} [\ell_{I_1, j}] - \min_{i \in \underline{N}} \ell_{i, j} \right) = f(A). \end{aligned}$$

b) leads to

$$0 = R_T(G) = \inf_A \bar{R}_T(A, G) \geq \inf_A f(A).$$

Observe that $f(A)$ depends on A through only the distribution of I_1 on \underline{N} denoted by $q = q(A)$ now, that is, $f(A) = f'(q)$ for proper f' . This dependence is continuous on the compact domain of q , hence the infimum can be replaced by minimum. Thus $\min_q f'(q) \leq 0$, that is, there exists a q such that for all $j \in \underline{M}$, $\mathbf{E} [\ell_{I_1, j}] = \min_{i \in \underline{N}} \ell_{i, j}$. This implies that the support of q contains only actions whose loss is not larger than the loss of any other action irrespectively of the choice of Nature's action. (Such an action is obviously non-dominated as shown by any $p \in \Delta_M$ supported on all outcomes.)

c) \Rightarrow d) Action i in c) is non-dominated, and any other action with loss vector distinct from ℓ_i is dominated (by i and any action with loss vector ℓ_i).

d) \Rightarrow a) For any action $i \in \underline{N}$, as in the proof of Lemma 14, consider the compact convex cell C_i in Δ_M . By Lemma 15 $\bigcup_{i \in \underline{N}} C_i = \Delta_M$. This and d) imply that there is an i with $C_i = \Delta_M$, that is, i is optimal for any outcome. So the algorithm that always plays i has zero regret for all outcome sequences and T . \square

Proof of Theorem 2 Case (1d). We know that $K \geq 2$ and G has no revealing action. Then for any A ,

$$\begin{aligned}\bar{R}_T(A, G) &\geq \sup_{j \in \underline{M}, J_1 = \dots = J_T = j} \mathbf{E} \left[\sum_{t=1}^T \ell_{I_t, J_t} - \min_{i \in \underline{N}} \sum_{t=1}^T \ell_{i, J_t} \right] \\ &\geq \frac{1}{M} \sum_{j=1}^M \mathbf{E} \left[\sum_{t=1}^T \ell_{I_t, j} - T \min_{i \in \underline{N}} \ell_{i, j} \right] \\ &= \frac{1}{M} \sum_{t=1}^T \mathbf{E} \left[\sum_{j=1}^M \ell_{I_t, j} \right] - \frac{T}{M} \sum_{j=1}^M \min_{i \in \underline{N}} \ell_{i, j} .\end{aligned}$$

Here I_t is a random variable usually depending on $J_{1:T-1}$, that is, on j through the outcomes. However, since G has no revealing action, now the distribution of I_t is independent of j , thus $\mathbf{E}[\sum_{j=1}^M \ell_{I_t, j}] \geq \min_{i \in \underline{N}} \sum_{j=1}^M \ell_{i, j}$ for each t , and we have

$$\bar{R}_T(A, G) \geq T \underbrace{\frac{1}{M} \left[\min_{i \in \underline{N}} \sum_{j=1}^M \ell_{i, j} - \sum_{j=1}^M \min_{i \in \underline{N}} \ell_{i, j} \right]}_c = cT ,$$

where $c > 0$ if $K \geq 2$ (because $c \geq 0$, and $c = 0$ would imply Lemma 3 c), thus also d)). Since c depends only on \mathbf{L} , $R_T(G) \geq cT = \Theta(T)$. \square

Proof of Lemma 15. By Definition 1, action i is dominated if and only if $C_i \subseteq \bigcup_{i': \ell_{i'} \neq \ell_i} C_{i'}$.

$C_i \subseteq \bigcup_{i': \ell_{i'} \neq \ell_i} C_{i'} \Rightarrow \text{int } C_i = \emptyset$: Since $\ell_{i'} \neq \ell_i \Rightarrow i \neq i'$, follows from (5).

$\text{int } C_i = \emptyset \Rightarrow \lambda(C_i) = 0$: Follows from convexity of C_i .

$\lambda(C_i) = 0 \Rightarrow C_i \subseteq \bigcup_{i': \ell_{i'} \neq \ell_i} C_{i'}$: indirect: if $p \in C_i$ is in the complementer of $\bigcup_{i': \ell_{i'} \neq \ell_i} C_{i'}$, that is open in Δ_M , then there is a neighborhood S of p in Δ_M disjoint from $\bigcup_{i': \ell_{i'} \neq \ell_i} C_{i'}$. Thus $S \subseteq \bigcup_{i': \ell_{i'} = \ell_i} C_{i'} = C_i$ due to (4), and $\lambda(C_i) \geq \lambda(S) > 0$, contradiction.

Since $\lambda(\bigcup_{i \in \mathcal{D}} C_i) \leq \sum_{i \in \mathcal{D}} \lambda(C_i) = 0$, thus from (4) $\lambda(\bigcup_{i \notin \mathcal{D}} C_i) \geq \lambda(\Delta_M)$, and $\lambda(\Delta_M \setminus \bigcup_{i \notin \mathcal{D}} C_i) = 0$. The latest set is open in Δ_M , so it must be empty, that is, $\bigcup_{i \notin \mathcal{D}} C_i = \Delta_M$. \square

Proof of Lemma 17. Clearly, the worst-case expected regret of A is at least its average regret:

$$\bar{R}_T(A, G) = \sup_{j_{1:T} \in \underline{M}^T} R_T(A, G) \geq \mathbf{E}_k[R_T(A, G)] = \mathbf{E}_k \left[\sum_{t=1}^T \ell_{I_t, J_t} - \min_{i \in \underline{N}} \sum_{t=1}^T \ell_{i, J_t} \right] ,$$

where the expectation on the right-hand side is taken with respect to both the random choices of the outcomes and the internal randomization of A . We lower bound the right-hand side switching expectation and minimum to get

$$\begin{aligned}\mathbf{E}_k \left[\sum_{t=1}^T \ell_{I_t, J_t} - \min_{i \in \underline{N}} \sum_{t=1}^T \ell_{i, J_t} \right] &\geq \sum_{t=1}^T \mathbf{E}_k \ell_{I_t, J_t} - \min_{i \in \underline{N}} \sum_{t=1}^T \mathbf{E}_k \ell_{i, J_t} \\ &= \sum_{t=1}^T \sum_{i=1}^N \mathbf{E}_k [\mathbb{I}(I_t = i) \ell_{i, J_t}] - \min_{i \in \underline{N}} \sum_{t=1}^T \langle \ell_i, p_k \rangle \\ &= \sum_{t=1}^T \sum_{i=1}^N \mathbf{E}_k \mathbb{I}(I_t = i) \mathbf{E}_k \ell_{i, J_t} - T \min_{i \in \underline{N}} \langle \ell_i, p_k \rangle\end{aligned}$$

$$\begin{aligned}
& \text{(by the independence of } I_t \text{ and } J_t) \\
&= \sum_{i=1}^N \langle \ell_i, p_k \rangle \sum_{t=1}^T \Pr_k[I_t = i] - T \min_{i \in \underline{N}} \langle \ell_i, p_k \rangle \\
&= \sum_{i=1}^N N_i^{(k)} \langle \ell_i, p_k \rangle - T \langle \ell_k, p_k \rangle \\
&= \sum_{\substack{i \in \underline{N} \\ i \neq k}} N_i^{(k)} \langle \ell_i - \ell_k, p_k \rangle .
\end{aligned} \tag{A.1}$$

(A.1) follows from the fact that action k is optimal under p_k . Clearly the term $i = k$ can be omitted in the last equality. \square

Proof of Lemma 18. We only prove the first inequality, the other one is symmetric. Assume first that A is deterministic, that is, $I_t : \Sigma^{t-1} \rightarrow \underline{N}$, and so $I_t(h_{1:t-1})$ denotes the choice of the algorithm at time step t , given that the (random) history of observations of length $t - 1$, $H_{1:t-1} = (H_1, \dots, H_{t-1})$ takes $h_{1:t-1} = (h_1, \dots, h_{t-1}) \in \Sigma^{t-1}$. (Note that this is a slightly different history definition than $\mathcal{H}_{1:t-1}$ defined in Section 5.1, as $H_{1:t-1}$ does not include the actions since their choices are determined by the feedback anyway. In general, $\mathcal{H}_{1:t-1}$ is equivalent to $H_{1:t-1} \cup (I_1, \dots, I_{t-1})$. Nevertheless, if it is assumed that the feedback symbol sets of actions are disjoint then $H_{1:t-1}$ and $\mathcal{H}_{1:t-1}$ are equivalent.) We denote by p_k^* the joint distribution of $H_{1:T-1}$ over Σ^{T-1} associated with p_k . (For games with only all-revealing actions, assuming $h_{i,j} = j$ in \mathbf{H} , p_k^* is the product distribution over the outcome sequences, that is, formally, $p_k^*(j_{1:T-1}) = \prod_{t=1}^{T-1} p_k(j_t)$.) We can bound the difference $N_2^{(2)} - N_2^{(1)}$ as

$$\begin{aligned}
N_i^{(2)} - N_i^{(1)} &= \sum_{t=1}^T (\Pr_2[I_t = i] - \Pr_1[I_t = i]) \\
&= \sum_{h_{1:T-1} \in \Sigma^{T-1}} \sum_{t=1}^T (\mathbb{I}(I_t(h_{1:t-1}) = i) p_2^*(h_{1:T-1}) - \mathbb{I}(I_t(h_{1:t-1}) = i) p_1^*(h_{1:T-1})) \\
&= \sum_{h_{1:T-1} \in \Sigma^{T-1}} (p_2^*(h_{1:T-1}) - p_1^*(h_{1:T-1})) \cdot \sum_{t=1}^T \mathbb{I}(I_t(h_{1:t-1}) = i) \\
&\leq T \sum_{\substack{h_{1:T-1} \in \Sigma^{T-1} \\ p_2^*(h_{1:T-1}) \geq p_1^*(h_{1:T-1})}} (p_2^*(h_{1:T-1}) - p_1^*(h_{1:T-1})) \\
&= \frac{T}{2} \|p_2^* - p_1^*\|_1 \\
&\leq T \sqrt{D(p_2^* \| p_1^*)/2} ,
\end{aligned} \tag{A.2}$$

where the last step is an application of Pinsker's inequality [24, Lemma 12.6.1] to distributions p_1^* and p_2^* . Using the chain rule for KL divergence [24, Theorem 2.5.3] we can write (with somewhat sloppy notation)

$$D(p_2^* \| p_1^*) = \sum_{t=1}^{T-1} D(p_2^*(h_t | h_{1:t-1}) \| p_1^*(h_t | h_{1:t-1})) ,$$

where the t^{th} conditional KL divergence term is

$$\sum_{h_{1:t-1} \in \Sigma^{t-1}} \Pr_2(H_{1:t-1} = h_{1:t-1}) \sum_{h_t \in \Sigma} \Pr_2(H_t = h_t \mid H_{1:t-1} = h_{1:t-1}) \ln \frac{\Pr_2(H_t = h_t \mid H_{1:t-1} = h_{1:t-1})}{\Pr_1(H_t = h_t \mid H_{1:t-1} = h_{1:t-1})} . \quad (\text{A.3})$$

Decompose this sum for the case $I_t(h_{1:t-1}) \notin \mathcal{R}$ and $I_t(h_{1:t-1}) \in \mathcal{R}$. In the first case, we play a none-revealing action, thus our observation $H_t = h_{I_t(h_{1:t-1}), J_t} = h_{I_t(h_{1:t-1}), 1}$ is a deterministic constant in both models 1 and 2, thus both $\Pr_1(\cdot \mid H_{1:t-1} = h_{1:t-1})$ and $\Pr_2(\cdot \mid H_{1:t-1} = h_{1:t-1})$ are degenerate and the KL divergence factor is 0. Otherwise, playing a revealing action, $H_t = h_{I_t(h_{1:t-1}), J_t}$ is the same deterministic function of J_t (which is independent of $H_{1:t-1}$) in both models 1 and 2, and so the inner sum in (A.3) is

$$\sum_{h_t \in \Sigma} \Pr_2[h_{I_t(h_{1:t-1}), J_t} = h_t] \ln \frac{\Pr_2[h_{I_t(h_{1:t-1}), J_t} = h_t]}{\Pr_1[h_{I_t(h_{1:t-1}), J_t} = h_t]} . \quad (\text{A.4})$$

Since $\Pr_k[h_{I_t(h_{1:t-1}), J_t} = h_t] = \sum_{j_t \in \underline{M}: h_{I_t(h_{1:t-1}), j_t} = h_t} p_k(j_t)$ ($k = 1, 2$), using the log sum inequality [24, Theorem 2.7.1]), (A.4) is upper bounded by

$$\sum_{h_t \in \Sigma} \sum_{j_t \in \underline{M}: h_{I_t(h_{1:t-1}), j_t} = h_t} p_2(j_t) \ln \frac{p_2(j_t)}{p_1(j_t)} = \sum_{j_t \in \underline{M}} p_2(j_t) \ln \frac{p_2(j_t)}{p_1(j_t)} = D(p_2 \parallel p_1) .$$

Hence, $D(p_2^* \parallel p_1^*)$ is upper bounded by

$$\sum_{t=1}^{T-1} \sum_{\substack{h_{1:t-1} \in \Sigma^{t-1} \\ I_t(h_{1:t-1}) \in \mathcal{R}}} \Pr_2(H_{1:t-1} = h_{1:t-1}) D(p_2 \parallel p_1) = D(p_2 \parallel p_1) \sum_{t=1}^{T-1} \sum_{i \in \mathcal{R}} \Pr_2[I_t = i] = D(p_2 \parallel p_1) N_{\text{rev}}^{(2, T-1)} ,$$

where $N_{\text{rev}}^{(k, T-1)} = \sum_{t=1}^{T-1} \Pr_k[I_t \in \mathcal{R}]$. This together with (A.2) gives $N_i^{(2)} - N_i^{(1)} \leq T \sqrt{D(p_2 \parallel p_1) N_{\text{rev}}^{(2, T-1)}} / 2$.

If A is random and its internal random “bits” are represented by a random value Z (which is independent of J_1, J_2, \dots), then $N_i^{(k)} = \mathbf{E}[\tilde{N}_i^{(k)}(Z)]$ for $\tilde{N}_i^{(k)}(Z) = \sum_{t=1}^T \Pr_k[I_t = i | Z]$. Also let $\tilde{N}_{\text{rev}}^{(k, T-1)}(Z) = \sum_{t=1}^{T-1} \Pr_k[I_t \in \mathcal{R} | Z]$. The proof above implies that for any fixed $z \in \text{Range}(Z)$,

$$\tilde{N}_i^{(2)}(z) - \tilde{N}_i^{(1)}(z) \leq T \sqrt{D(p_2 \parallel p_1) \tilde{N}_{\text{rev}}^{(2, T-1)}(z)} / 2 ,$$

and thus, using also Jensen’s inequality,

$$\begin{aligned} N_i^{(2)} - N_i^{(1)} &= \mathbf{E}[\tilde{N}_i^{(2)}(Z) - \tilde{N}_i^{(1)}(Z)] \\ &\leq \mathbf{E}\left[T \sqrt{D(p_2 \parallel p_1) \tilde{N}_{\text{rev}}^{(2, T-1)}(Z)} / 2\right] \\ &\leq T \sqrt{D(p_2 \parallel p_1) \mathbf{E}[\tilde{N}_{\text{rev}}^{(2, T-1)}(Z)]} / 2 = T \sqrt{D(p_2 \parallel p_1) N_{\text{rev}}^{(2, T-1)}} / 2 , \end{aligned}$$

that is clearly upper bounded by $T \sqrt{D(p_2 \parallel p_1) N_{\text{rev}}^{(2)}} / 2$ yielding the statement of the lemma. \square

Proof of Lemma 20. We prove the lemma by showing that for every algorithm A on game G there exists an algorithm A' on G' such that for any outcome sequence, $R_T(A', G') \leq R_T(A, G)$ and vice versa. Recall that the minimax regret of a game is

$$R_T(G) = \inf_A \sup_{J_{1:T} \in \underline{M}^T} R_T(A, G) ,$$

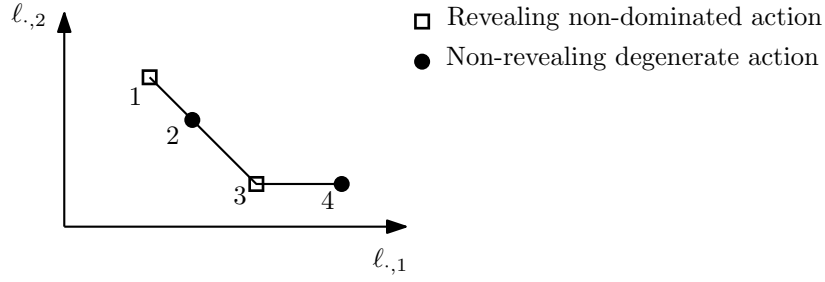


Figure A.10: Degenerate non-revealing actions on the chain. The loss vector of action 2 is a convex combination of that of action 1 and 3. On the other hand, the loss vector of action 4 is component-wise lower bounded by that of action 3.

where

$$R_T(A, G) = \mathbf{E} \left[\sum_{t=1}^T \ell_{I_t, J_t} - \min_{i \in N} \sum_{t=1}^T \ell_{i, J_t} \right],$$

First we observe that the term $\mathbf{E}[\min_{i \in N} \sum_{t=1}^T \ell_{i, J_t}]$ does not change by removing degenerate actions. Indeed, by the definition of degenerate action, if the minimum is given by a degenerate action then there exists a non-degenerate action with the same cumulative loss. It follows that we only have to deal with the term $\mathbf{E}[\sum_{t=1}^T \ell_{I_t, J_t}]$.

1. Let A' be an algorithm on G' . We define the algorithm A on G by choosing the same actions as A' at every time step. Since the action set of G is a superset of that of G' , this construction results in a well defined algorithm on G , and trivially has the same expected loss as A' .
2. Let A be an algorithm on G . From the definition of degenerate actions, we know that for every degenerate action i , there are two possibilities:
 - (a) There exists a non-degenerate action i_1 such that ℓ_i is component-wise lower bounded by ℓ_{i_1} .
 - (b) There are two non-degenerate actions i_1 and i_2 such that ℓ_i is a convex combination of ℓ_{i_1} and ℓ_{i_2} , that is, $\ell_i = \alpha_i \ell_{i_1} + (1 - \alpha_i) \ell_{i_2}$ for some $\alpha_i \in (0, 1)$.

An illustration of these cases can be found in Figure A.10. We construct A' the following way. At every time step t , if I_t^A (the action that algorithm A would take) is non-degenerate then let $I_t^{A'} = I_t^A$. If $I_t^A = i$ is a degenerate action of the first kind, let $I_t^{A'}$ be i_1 . If $I_t^A = i$ is a degenerate action of the second kind then let $I_t^{A'}$ be i_1 with probability α_i and i_2 with probability $1 - \alpha_i$. Recall that G is non-degenerate, so i has to be a non-revealing action. However, i_1 and/or i_2 might be revealing ones. To handle this, A' is defined to map the observation sequence, before using it as the argument of I_t , replacing the feedbacks corresponding to degenerate action i by $h_{i,1} = h_{i,2}$. That is, intuitively, A' “pretends” that the feedbacks at such time steps are irrelevant. It is clear that the expected loss of A' in every time step is less than or equal to the expected loss of A , concluding the proof. □

Proof of Theorem 13 for adversarial nature

For the proof, we start with a lemma, which ensures the existence of a pair i_1, i_2 of actions and an outcome distribution p with M atoms such that both i_1 and i_2 are optimal under p .

Lemma 22. *Let $G = (\mathbf{L}, \mathbf{H})$ be any finite non-trivial game with N actions and $M \geq 2$ outcomes. Then there exists $p \in \Delta_M$ satisfying both of the following properties:*

(a) All coordinates of p are positive.

(b) There exist actions $i_1, i_2 \in \underline{N}$ such that $\ell_{i_1} \neq \ell_{i_2}$ and for all $i \in \underline{N}$,

$$\langle \ell_{i_1}, p \rangle = \langle \ell_{i_2}, p \rangle \leq \langle \ell_i, p \rangle .$$

Proof of Lemma 22. Note that distributions p with positive coordinates form the interior of Δ_M ($\text{int } \Delta_M$). For any action $i \in \underline{N}$, as in the proof of Lemma 14, consider the compact convex cell C_i in Δ_M , whose union is Δ_M (see (4)). Let p_1 be any point in the interior of Δ_M . By (4), there is a cell C_{i_1} containing p_1 . If $C_{i_1} = \Delta_M$ held then action i_1 would satisfy Lemma 3 c), thus also d), and the game would be trivial. So there must be a point, say p_2 , in $\Delta_M \setminus C_{i_1}$. The intersection of the closed segment $\overline{p_1 p_2}$ and C_{i_1} is closed and convex, thus it is a closed subsegment $\overline{p_1 p}$ for some $p \in C_{i_1}$ ($p \neq p_2$). $p_1 \in \text{int } \Delta_M$ and the convexity of Δ_M imply $p \in \text{int } \Delta_M$. Since the open segment $\overline{p p_2}$ has to be covered by $\bigcup_{i': C_{i'} \neq C_{i_1}} C_{i'}$, that is a closed set, $p \in \bigcup_{i': C_{i'} \neq C_{i_1}} C_{i'}$ must also hold, that is, $p \in C_{i_2}$ for some $C_{i_2} \neq C_{i_1}$ (requiring $\ell_{i_1} \neq \ell_{i_2}$). Hence p satisfies both (a) and (b). \square

Proof of Theorem 13. When $M = 1$, G is always trivial, thus we assume that $M \geq 2$. Without loss of generality we may assume that all the actions are all-revealing.

Let $p \in \Delta_M$ be a distribution of the outcomes that satisfies conditions (a) and (b) of Lemma 22. By renaming actions we can assume without loss of generality that $\ell_1 \neq \ell_2$ and actions 1 and 2 are optimal under p , that is,

$$\langle \ell_1, p \rangle = \langle \ell_2, p \rangle \leq \langle \ell_i, p \rangle \quad (\text{A.5})$$

for any $i \in \underline{N}$.

Fix any learning algorithm A . We use randomization replacing the outcomes by a sequence J_1, J_2, \dots, J_T of random variables i.i.d. according to p , and independent of the internal randomization of A . Clearly, as in the proof of Lemma 17, the worst-case expected regret of A is at least its average regret:

$$\bar{R}_T(A, G) \geq \mathbf{E}[R_T(A, G)] = \mathbf{E} \left[\sum_{t=1}^T \ell_{I_t, J_t} - \min_{i \in \underline{N}} \sum_{t=1}^T \ell_{i, J_t} \right] = \mathbf{E} \left[\sum_{t=1}^T \mathbf{E}[\ell_{I_t, J_t} | I_t] - \min_{i \in \underline{N}} \sum_{t=1}^T \ell_{i, J_t} \right]. \quad (\text{A.6})$$

Here, in the last two expressions, the expectation is with respect to both the internal randomization of A and the random choice of J_1, J_2, \dots, J_T . Now, since J_t is independent of I_t , we see that $\mathbf{E}[\ell_{I_t, J_t} | I_t] = \langle \ell_{I_t}, p \rangle$. By (A.5), we have $\langle \ell_{I_t}, p \rangle \geq \langle \ell_1, p \rangle = \langle \ell_2, p \rangle$. Therefore (upper bounding also the minimum),

$$\begin{aligned} \sum_{t=1}^T \mathbf{E}[\ell_{I_t, J_t} | I_t] - \min_{i \in \underline{N}} \sum_{t=1}^T \ell_{i, J_t} &= \sum_{t=1}^T \langle \ell_{I_t}, p \rangle - \min_{i \in \underline{N}} \sum_{t=1}^T \ell_{i, J_t} \\ &\geq \sum_{t=1}^T \langle \ell_1, p \rangle - \min_{i=1,2} \sum_{t=1}^T \ell_{i, J_t} \\ &= \max_{i=1,2} \sum_{t=1}^T (\langle \ell_1, p \rangle - \ell_{i, J_t}) . \end{aligned} \quad (\text{A.7})$$

Using the identity $\max\{a, b\} = \frac{1}{2}(a + b + |a - b|)$, the latest expression is

$$\begin{aligned} & \frac{1}{2} \left[\sum_{t=1}^T (\langle \ell_1, p \rangle - \ell_{1,J_t}) + \sum_{t=1}^T (\langle \ell_1, p \rangle - \ell_{2,J_t}) + \left| \sum_{t=1}^T (\langle \ell_1, p \rangle - \ell_{1,J_t}) - \sum_{t=1}^T (\langle \ell_1, p \rangle - \ell_{2,J_t}) \right| \right] \\ &= \frac{1}{2} \sum_{t=1}^T (\langle \ell_1, p \rangle - \ell_{1,J_t} + \langle \ell_2, p \rangle - \ell_{2,J_t}) + \frac{1}{2} \left| \sum_{t=1}^T (\ell_{2,J_t} - \ell_{1,J_t}) \right|, \end{aligned}$$

where (A.5) was used in the first term. The expectation of the first term vanishes since $\mathbf{E}[\ell_{i,J_t}] = \langle \ell_i, p \rangle$. Let $X_t = \ell_{2,J_t} - \ell_{1,J_t}$. We see that X_1, X_2, \dots, X_T are i.i.d. random variables with mean $\mathbf{E}[X_t] = 0$. Therefore,

$$\mathbf{E} \left[\max_{i=1,2} \sum_{t=1}^T (\langle \ell_i, p \rangle - \ell_{i,J_t}) \right] = \frac{1}{2} \mathbf{E} \left[\sum_{t=1}^T X_t \right] \geq c \sqrt{T}, \quad (\text{A.8})$$

where the last inequality follows from Theorem 23 stated below and the constant c depends only on ℓ_1, ℓ_2 , and p . For the theorem to yield $c > 0$, it is important to note that the distribution of X_t has finite support and with positive probability $X_t \neq 0$ since $\ell_1 \neq \ell_2$ and all coordinates of p are positive. Hence, both $\mathbf{E}[X_t^2]$ and $\mathbf{E}[X_t^4]$ are finite and positive.

Now, putting together (A.6), (A.7), and (A.8) gives the desired lower bound $\bar{R}_T(A, G) \geq c \sqrt{T}$. Since c depends only on \mathbf{L} , also $R_T(G) \geq c \sqrt{T}$. \square

The following theorem is a variant of Khinchine's inequality (see e.g. [20, Lemma A.9]) for asymmetric random variables. The idea of the proof is the same as there and originally comes from Littlewood [25].

Theorem 23 (Khinchine's inequality for asymmetric random variables). *Let X_1, X_2, \dots, X_T be i.i.d. random variables with mean $\mathbf{E}[X_t] = 0$, finite variance $\mathbf{E}[X_t^2] = \text{Var}(X_t) = \sigma^2$, and finite fourth moment $\mathbf{E}[X_t^4] = \mu_4$. Then,*

$$\mathbf{E} \left[\sum_{t=1}^T X_t \right] \geq \frac{\sigma^3}{\sqrt{3\mu_4}} \sqrt{T}.$$

Proof. [26, Lemma A.4] implies that for any random variable Z with finite fourth moment

$$\mathbf{E}|Z| \geq \frac{(\mathbf{E}[Z^2])^{3/2}}{(\mathbf{E}[Z^4])^{1/2}}.$$

Applying this inequality to $Z = \sum_{t=1}^T X_t$ we get

$$\mathbf{E} \left[\sum_{t=1}^T X_t \right] \geq \frac{T^{3/2} \sigma^3}{T \sqrt{3\mu_4}} = \frac{\sigma^3}{\sqrt{3\mu_4}} \sqrt{T},$$

that follows from

$$\mathbf{E}[Z^2] = \mathbf{E} \left[\left(\sum_{t=1}^T X_t \right)^2 \right] = \sum_{t=1}^T \mathbf{E}[X_t^2] = T \sigma^2$$

and

$$\mathbf{E}[Z^4] = \mathbf{E} \left[\left(\sum_{t=1}^T X_t \right)^4 \right] = \sum_{t=1}^T \mathbf{E}[X_t^4] + 6 \sum_{1 \leq s < t \leq T} \mathbf{E}[X_s^2] \mathbf{E}[X_t^2] = T \mu_4 + 3T(T-1) \sigma^4 \leq 3T^2 \mu_4,$$

where we have used the independence of X_t 's and $\mathbf{E}[X_t] = 0$ which ensure that mixed terms $\mathbf{E}[X_t X_s]$, $\mathbf{E}[X_t X_s^3]$, etc. vanish. We also used that $\sigma^4 = \mathbf{E}[X_t^2]^2 \leq \mathbf{E}[X_t^4] = \mu_4$. \square

References

- [1] Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Toward a classification of finite partial-monitoring games. In *Proceedings of Algorithmic Learning Theory (ALT 2010), Canberra, Australia, September 6–8, 2010*, 2003.
- [2] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.
- [3] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 1(55):119–139, 1997.
- [4] Nicolò Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.
- [5] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.
- [6] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. Online optimization in X-armed bandits. In *Advances in Neural Information Processing Systems 21 (NIPS)*, pages 201–208, 2009.
- [7] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th annual ACM Symposium on Theory of Computing (STOC 2008)*, pages 681–690. ACM, 2008.
- [8] David Helmbold and Sandra Panizza. Some label efficient learning results. In *Proceedings of the 10th Annual Conference on Computational Learning Theory (COLT 1997)*, pages 218–230. ACM, 1997.
- [9] Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Minimizing regret with label efficient prediction. *IEEE Transactions on Information Theory*, 51(6):2152–2162, June 2005.
- [10] Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *Proceedings of 44th Annual IEEE Symposium on Foundations of Computer Science 2003 (FOCS 2003)*, pages 594–605. IEEE, 2003.
- [11] Avrim Blum and Jason D. Hartline. Near-optimal online auctions. In *Proceedings of the 16th Annual ACM-SIAM symposium on Discrete Algorithms (SODA 2005)*, pages 1156–1163. Society for Industrial and Applied Mathematics, 2005.
- [12] Alekh Agarwal, Peter Bartlett, and Max Dama. Optimal allocation strategies for the dark pool problem. In *13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010), May 12-15, 2010, Chia Laguna Resort, Sardinia, Italy*, 2010.
- [13] David P. Helmbold, Nicholas Littlestone, and Philip M. Long. Apple tasting. *Information and Computation*, 161(2):85–139, 2000.
- [14] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of Twentieth International Conference on Machine Learning (ICML 2003)*, 2003.
- [15] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. Technical Report: CMU-CS-03-110, 2003. Available at: <http://reports-archive.adm.cs.cmu.edu/anon/anon/usr0/ftp/2003/CMU-CS-03-110.pdf>.
- [16] Jacob Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, pages 263–273. Citeseer, 2008.
- [17] Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the 16th annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2005)*, page 394. Society for Industrial and Applied Mathematics, 2005.
- [18] James Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.
- [19] Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *Proceedings of the 14th Annual Conference on Computational Learning Theory (COLT 2001)*, pages 208–223. Springer-Verlag, 2001.
- [20] Gábor Lugosi and Nicolò Cesa-Bianchi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- [21] Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.
- [22] Nicolò Cesa-Bianchi, Gábor Lugosi, and Gilles Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31(3):562–580, 2006.
- [23] Gábor Lugosi, Shie Mannor, and Gilles Stoltz. Strategies for prediction under imperfect monitoring. *Mathematics of Operations Research*, 33(3):513–528, 2008.
- [24] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, New York, second edition, 2006.
- [25] John E. Littlewood. On bounded bilinear forms in an infinite number of variables. *The Quarterly Journal of Mathematics*, 1: 164–174, 1930.
- [26] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Applications of Mathematics: Stochastic Modelling and Applied Probability. Springer-Verlag New York, 1996.